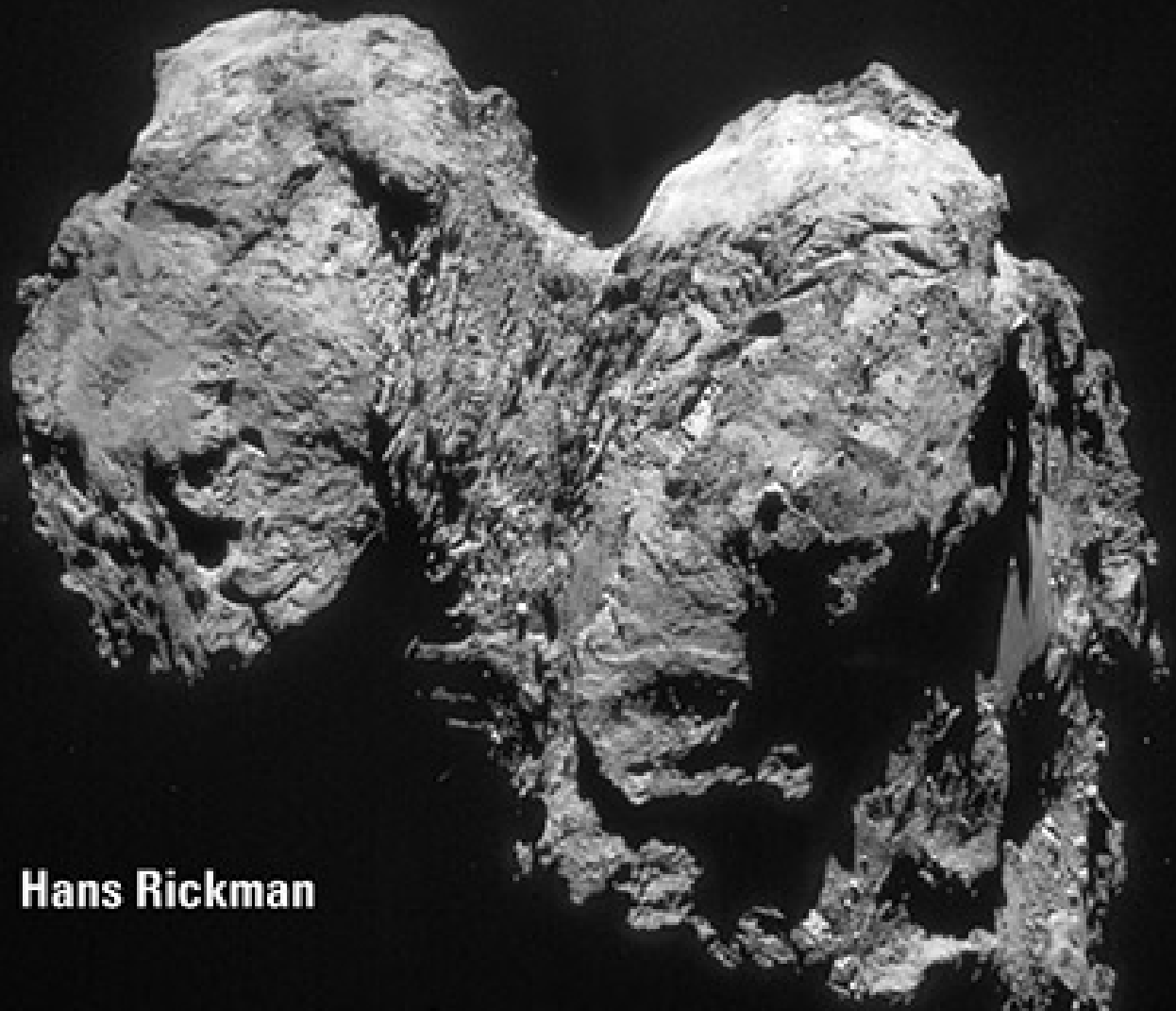
Advances in Planetary Science – Vol. 2

ORIGIN AND EVOLUTION OF COMETS

Ten Years after the Nice Model and One Year after Rosetta



Hans Rickman

World Scientific

Advances in Planetary Science – Vol. 2

ORIGIN AND EVOLUTION OF COMETS

Ten Years after the Nice Model and One Year after Rosetta

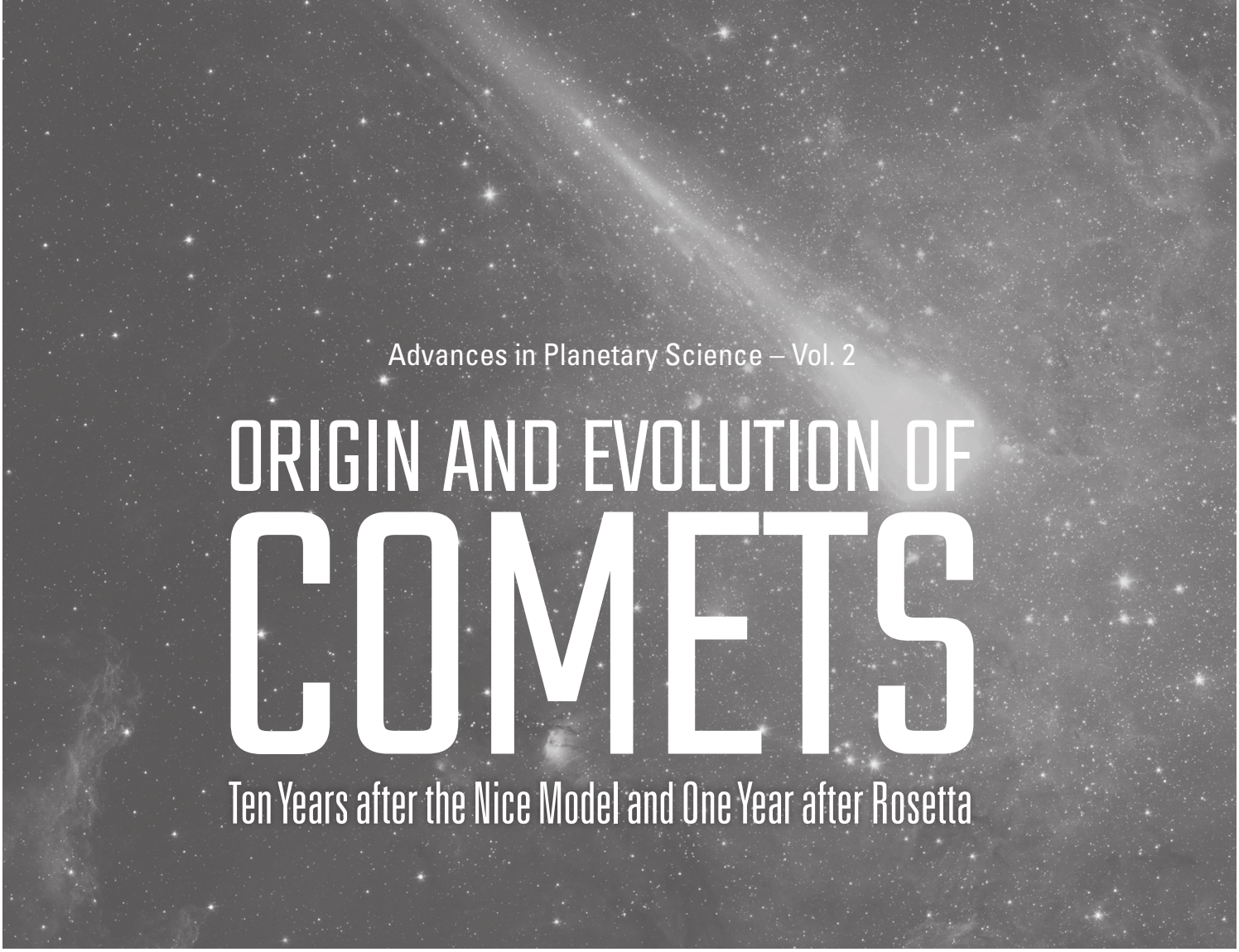
Advances in Planetary Science

Series Editor: Wing-Huen Ip *(National Central University, Taiwan)*

Published

Vol. 2 *Origin and Evolution of Comets:*
Ten Years after the Nice Model and One Year after Rosetta
by Hans Rickman

Vol. 1 *Nuclear Planetary Science: Planetary Science Based on Gamma-Ray, Neutron and X-Ray Spectroscopy*
by Nobuyuki Hasebe, Kyeong Ja Kim, Eido Shibamura and Kunitomo Sakurai



Advances in Planetary Science – Vol. 2

ORIGIN AND EVOLUTION OF COMETS

Ten Years after the Nice Model and One Year after Rosetta

Hans Rickman

Uppsala University, Sweden & PAS Space Research Center, Poland

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI • TOKYO

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data
Names: Rickman, H. (Hans), author.
Title: Origin and evolution of comets : ten years after the Nice model and one year after Rosetta / Hans Rickman (Uppsala University, Sweden & Polish Academy of Sciences, Poland).
Description: Singapore ; Hackensack, NJ : World Scientific, [2017] | Series: Advances in planetary science ; volume 2 | Includes bibliographical references and index.
Identifiers: LCCN 2017010483| ISBN 9789813222571 (hard cover ; alk. paper) | ISBN 9813222573 (hard cover ; alk. paper)
Subjects: LCSH: Comets--History.
Classification: LCC QB721 .R634 2017 | DDC 523.6/6--dc23
LC record available at https://lccn.loc.gov/2017010483

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Desk Editor: Ng Kah Fee

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore

Preface

In August 2005, the international research community concerning the small bodies of the solar system gathered for the triennial *Asteroids, Comets, Meteors* meeting in the small seaside resort of Buzios north of Rio de Janeiro, Brazil. I remember with pleasure the nice Brazilian winter and the interesting discussions we had around the hot topics of those days. The one thing that stands out in my recollections is the first news of the Nice Model — the term had not yet been coined, but it was immediately clear that this was extremely important. Something else, which did not attract so much attention while being a potential time bomb for the future, was the fact that ESA's Rosetta probe was on its way toward its asteroidal and cometary targets since more than a year.

Of course, the Nice Model is principally about planets and, at least for myself, its repercussions for cometary science took some time to grasp. But today it is practically impossible to discuss the origin and early evolution of comets without reference to the Nice Model. In addition, after all, according to this scenario, comets were the real arbiters of the planetary system in the early days, when they governed the dynamics of the giant planets by their collective gravitational effects. We call these icy planetesimals instead of comets, but it is the same objects that nowadays populate the Oort Cloud, the Edgeworth–Kuiper Belt and the Scattered Disk.

It is worth emphasizing that the central theme of the Nice Model remains unchanged, but there are many variations around this theme with different implications for the comets. To account for this was one of my goals when writing this book.

Concerning Rosetta, I was involved from the very beginning, and I played a part in the OSIRIS camera team largely by procuring the filters. I hence had a personal involvement and lived through the mission with my own personal hopes and worries. The mission concept was novel and daring, and people like myself who were not born optimists had a hard time realizing what was actually to come. But, eventually, the time bomb did burst, and it felt like a two-year long roller coaster ride. So many discoveries, so many stunning images, thanks to the sterling work by all involved! This is an experience that may rightly be called a privilege to have been part of.

Now, after the end of the mission, it is time to assess what we have learned. I am convinced that, one day, this will be done and our knowledge about comets will take a big step forward. However, I cannot reach such an ambitious goal in this book. For the moment, the big picture remains too dim, at least to my mind. In all modesty, I can only present some preliminary ideas and impressions that may be right or wrong. Hopefully, together with those of my international colleagues, they will be food for thought and discussion, awaiting the final breakthrough.

In the meantime, this book can be used as an introduction to cometary science from the perspective of origins and evolution. It is not intended to be a full account of the whole subject, and even important results may occasionally have been left out in the interest of brevity, to avoid boring the reader with too much details. While trying to be fair, I did not strive to give equal weight to all ideas or to downplay the natural bias toward my own opinions. The interested reader will find enough references to published literature in order to pursue any specific quest that may arise.

Let me finally express my heart-felt thanks to Tomek Wiśniowski for invaluable help with many of the illustrations, as well as Mike A'Hearn, Björn Davidsson, Gerhard Hahn, Alessandro Morbidelli

and Giovanni Valsecchi for helpful discussions. Above all, this book could hardly exist without the patience of my wife Bożenna, who, practically without complaint, endured a year of hardship when I was busy writing all the time.

Arboga, 3 February, 2017,
Hans Rickman

This page intentionally left blank

Contents

Chapter 1

Introduction

In most parts of the world, it would be difficult to find people of age, who have no idea what a comet is. However, those ideas are generally different from the concepts that scientists have in mind when using the word comet. In fact, the scientific definition of a comet is a non-trivial issue, which we had better tackle before describing their physical properties and how they originated and evolved.

1.1. What is a Comet?

There are two interpretations of the word. Closest to the layman's impression is the one about the phenomenon observed on the sky, and according to this, a comet is a diffuse object whose technical term is *coma* (usually, including a bright spot called the *central condensation*), from which a *tail* may extend. This object is in orbit around the Sun. Referring to the orbit is often necessary to distinguish the comets from Galactic nebulae and external galaxies. In 1771, the first catalogue of such diffuse nebulae (the famous Messier catalogue) was in fact produced by Charles Messier in order to avoid wasting time on these objects when hunting for comets, because one sort of diffuse object was often difficult to distinguish from the other.

The second interpretation refers to a physical object belonging to the solar system. In 2006, the International Astronomical Union (IAU) at its 26th General Assembly adopted a classification of solar system objects in terms of *planets*, *dwarf planets* and *small bodies*.

In this definition, comets are counted with the small bodies together with, for instance, the asteroids. Here, the word comet means the solid object, orbiting around the Sun, which gives rise to the diffuse phenomena mentioned above. The mechanism whereby this occurs is another matter.

Generally, it is a question of ice sublimation due to heating by absorption of sunlight, which leads to an outflow of gas and dust into space. This outflowing material is seen as the coma with a possible, more or less anti-sunward extension, called the tail. The solid object, from which the coma and tail would emanate, is called the *nucleus*, and thus, the word comet is used as a synonym to the nucleus. It is fair to say that this usage dominates in recent scientific literature, and the same practice will be followed here. This is natural, because the concepts of origin and evolution always refer to the nucleus. The coma and tail evolve very rapidly and they typically come and go, as the comet moves around the Sun. But the nucleus persists and typically evolves on much more significant time scales encompassing many orbits.

The physical definition of a comet, as briefly sketched above, is in fact a bit ambiguous. If we consider ice sublimation as the cause of the outflow, it is obvious that this is strongly temperature dependent. Hence, it will depend on the distance between the object and the Sun. It is indeed a well-known fact that comets develop their comae and tails essentially in the innermost parts of their orbits. Thus, when comets are observed far from the Sun, they may appear starlike, and we may actually see the bare nucleus. Comets do not have to produce the diffuse cloud all the time, but reliable observations of such activity on at least one occasion are required. Thus, official status as a comet is recognized only after such observations, but it is not withdrawn if an established activity ceases.

Hence, the comet is in reality an object that has the potential to develop a coma and a tail under the right circumstances — essentially, when it comes close enough to the Sun. This means that the object has to contain ice but also that the ice must be found near the surface, so that sublimation may lead to an outflow of gas

and dust. Moreover, the perihelion distance must be small enough for this to happen. One can easily see that this is problematic. If the orbit is perturbed because of a close approach to Jupiter, the perihelion distance can change appreciably. An object may thus be called a comet, if it is discovered before an increase of the perihelion distance but not if the discovery happens afterwards. In addition, even slight modifications of the surface layers without any orbital change may imply that the gas production subsides or resumes, and hence, what is essentially the same object may or may not be called a comet depending on when it is observed — if not, it would probably be called an asteroid.

The only reasonable solution to these problems would be to call an icy object a comet, even if its perihelion distance is too large, or the ice is too deeply buried beneath the surface. The essential property of a comet would hence be its ice content: comets are icy, while asteroids are rocky or metallic. This definition is attractive in theory, because the criterion used is intrinsic and more or less quantifiable. Moreover, it helps to convey an important message, namely, that the small bodies of the solar system belong together, even though there is a range of chemical compositions depending on their formation temperatures. Comets and asteroids are not fundamentally different — they are different incarnations of the small body population, representing objects that were formed at different distances from the Sun and thus have different ice content. However, the definition is not useful in practice as long as we cannot measure the ice content by probing the interior of the objects, and we therefore have to require observed cometary activity as an objective criterion when distinguishing comets from asteroids.

In any case, it is clear that we have to be open minded about the objects to discuss. The border between comets and asteroids is somewhat fluent, and there may be transitional objects that are difficult to classify. When limiting ourselves to "real" comets that have exhibited comae or tails, we must recognize that they have siblings that sometimes need to be discussed in the same context.

1.1.1. *The comet nucleus*

Let us now pay some more attention to what a comet nucleus is thought to be, as an introduction to all the recent findings to be described below. As mentioned, the starting point is the gas and dust forming the comae and tails in comets. In the early 20th century, the old concept of a solid object within this cloud was no longer a dominating idea. Comets had been seen to split and disappear, and prominent meteor streams had been shown to trace the orbits of well known comets. It thus seemed natural to imagine a comet as nothing but a concentration of grains moving together in space. There was also a theory that claimed to show, how such comets could be formed by interstellar material captured by the gravity of the Sun, as it travels through the denser regions of interstellar space.

However, computations had shown that the observed comets do not show a tendency to arrive along hyperbolic orbits. Moreover, comets had been found to approach Jupiter closely without being dispersed and losing their identity, as one would expect from large clouds without much internal gravity. It thus seems, in retrospect, that there was no physical basis for the picture of comets as loose clouds. However, this was clear to some but not to all.

One problem was how to explain the origin of the coma by ice evaporation. In the 1940s there was little information about the chemical composition of the coma, but some radicals had been identified in comet spectra and shown to provide much of the light that is observed. In 1948, the Belgian astronomer Pol Swings proposed that these radicals were produced by the release and dissociation of ices made of polyatomic molecules. One obvious way to release these molecules was proposed by Fred Whipple (1950) in the paper that introduced the modern concept of a comet nucleus, namely, a solid body consisting of an *icy conglomerate* composed of ices and refractories in an intimate mix. Sublimation of the H_2O-dominated ice in the solar heat would release the parent molecules, from which the radicals emanate.

Whipple's paper dealt with one particular comet. This is *Encke's comet*, which was known since the early 19th century. With an orbital period of only 3.3 years it had been observed on many returns,

and scientists had noticed that each of these returns occurred a little too early, compared to the best predictions that could be made by integrating the orbit from the preceding apparitions. This *nongravitational effect* needed an explanation, and Whipple's solid nucleus offered a good explanation using the same concept as Bessel (1836) had used. This was a jet force acting on the nucleus due to the asymmetric outflow of material feeding the coma. Since observations of Encke and other comets indicated the outflow to occur mainly in the solar direction, this would mainly accelerate the nucleus in the radial direction outward from the Sun, and Bessel focused on this aspect. In Whipple's model, the asymmetry followed directly from the fact that the heating of the ice is strongest at the subsolar point, but a thermal lag due to the rotation of the nucleus could in principle add a transverse component to the radial acceleration. In principle, the latter could act persistently over time, if the rotation is markedly prograde or retrograde, and the effect would then be either too late or too early arrival at perihelion.

While Whipple's theory appeared to offer a good foundation for understanding the behavior of comets and thus seemed clearly preferable compared to its competitors, one problem would remain for decades. To explain the observed amounts of material in cometary comae, a km-sized nucleus was generally enough. The problem was that such a small object is very difficult to detect at the typical distances of observed comets, and the long-lasting absence of any clear observational verification of Whipple's nucleus caused some lingering skepticism by the proponents of alternative theories.

1.2. Comet Designations

In both media reports, popular descriptions and scientific literature, comets are referred to by names and designations. These are not always consistent and may appear confusing, so a brief guide may be helpful.

Comet orbits span an enormous range of revolution periods from just a few years to millions of years. Thus, comets can be subdivided into two categories: the single-apparition comets and the returning comets. The former often have so long orbital periods that,

essentially, astronomers have only had a single occasion to observe them in connection with one perihelion passage. The latter, on the other hand, have periods short enough to present at least two such occasions. Those categories have traditionally been referred to as long-period versus short-period comets (see Sec. 1.4), and the limit has been placed at orbital period $P = 200$ years.

The present designation system dates back to a resolution passed by the IAU 22nd General Assembly in 1994.[1] Here, the returning category is referred to as *periodic comets*. These are defined to have revolution periods of less than 200 years or confirmed observations at more than one perihelion passage.[2] Upon discovery, all previously unknown comets get a designation, consisting of the year of discovery followed by an upper-case letter denoting the halfmonth in question and a numeral indicating the sequential order of this discovery announcement within the relevant halfmonth. Before this, a letter is applied, which indicates the category of the comet: "P/" denotes a periodic comet, and "C/" denotes a comet that is not periodic.

In 1999, the IAU Minor Planet Center decided to call single-apparition comets periodic only when their orbital periods are less than 30 years. Meanwhile, there is another way to designate returning comets, which is independent of the orbital period. This is a permanent, serial number followed by the letter "P", and it is assigned to comets that have been observed to return or have had their periodicity established otherwise. The name of the comet may be added, separated by a slash. The list of such comets is basically chronologic, starting with 1P/Halley. As of 1 January 2016, there were 330 comets listed as periodic in this way, and on 1 January 2017 this number had grown to 347.

A category of special interest for the evolutionary aspect of comets is those that have been deemed not to exist any more. In

[1] www.minorplanetcenter.net/iau/lists/CometResolution.html

[2] To date, there is only one comet of the second kind, namely, comet 153P/Ikeya-Zhang with perihelia in 1661 and 2002. The 1661 observations were carried out by Jan Heweliusz, while the independent discoveries in 2002 were made by Kaoru Ikeya and Zhang Daqing.

most cases, the reason is that they have not been found in spite of deep exposures of the sky area where they would certainly have been according to the ephemerides. For these comets, the letter P is replaced by D in the designation. The list of comets with permanent numbers currently hosts eight such members, the most famous of which is 3D/Biela.

In addition, there are single-apparition comets with "D/" designations. Among these, the most notable is the famous, very special comet D/1993 F2 (Shoemaker-Levy 9), which was discovered in a jovicentric orbit and collided with Jupiter in 1994. This is the only comet that is definitely deceased, even though the letter D generally stands for 'dead' or 'disappeared'. In other cases, one has to consider the possibility that the comet will reappear. A case in point is that the record of single-apparition comets used to contain two members with designations D/1783 W1 (Pigott) and D/1819 W1 (Blanpain), which were long lost comets with orbits known to be of short period and were thought to have disappeared. However, both were recently rediscovered and are now known as 226P/Pigott-LINEAR-Kowalski and 289P/Blanpain, respectively.

Comets are often referred to by names. For some of them, this is almost indispensible since the names are so deeply rooted in people's minds. For instance, it would be strange to discuss comet 1P without adding the name Halley or D/1993 F2 without clarifying that this is comet Shoemaker-Levy 9. However, in contrast to the designations, names are never unique. For instance, even though it seems quite unlikely, it cannot be excluded that one day somebody named Halley will discover a bright comet, and people will be tempted to call this "Halley's comet". For this reason, names are treated as of secondary importance in the official designation system.

There are many idiosyncracies and complications associated with comet names and designations, which cannot be covered here.[3] However, a few items are worth noting. The difficulty of

[3] An account may be found in the IAU-endorsed comet naming guidelines of 2003; see http://www.ss.astro.umd.edu/IAU/csbn/cnames.shtml.

distinguishing between comets and asteroids has already been dealt with, and one consequence is that many comets were designated as asteroids before their cometary nature was recognized. In such cases, the asteroidal designations are retained in conjunction with a cometary marker. One example is comet P/1999 DN_3, now known as 183P/Korlević-Jurić.

The names given to comets usually refer to their discoverers. Traditionally, these have been persons, but in recent times most comets are discovered due to team efforts within search programs or using space telescopes. Thus, many comets are named for the teams or space projects, like comets C/1999 S4 (LINEAR) or C/1997 B3 (SOHO). There are also a few periodic comets, which are named not after the discoverers but after the persons who investigated their orbits or established their periodic nature (1P/Halley, 2P/Encke and 27P/Crommelin).

At the time in 1994, when the current comet designation system was endorsed, one issue was especially controversial. In the 20th century, a practice had developed, whereby different periodic comets with the same discoverer were distinguished by adding a sequential number after the discoverer's name. With the introduction of the unique, permanent numbers, this secondary numbering was deemed redundant and thus abolished. However, in many cases those numbers have become an integral part of the name in the minds of comet scientists (and also, the general public as in the case of Shoemaker-Levy 9), and so they are still in frequent use. For instance, the target comet of the NASA Deep Impact mission is usually called 9P/Tempel 1, and the follow-up (EPOXI) target is usually called 103P/Hartley 2, even though the official names are 9P/Tempel and 103P/Hartley.

The last point to note here is that some objects have found their place in the permanent records of both comets and asteroids. Thus, they are recognized to have dual status. This group is heterogeneous as to both orbits and physical properties. The first object was asteroid (2060) Chiron — discovered in 1977 as the first among the so-called *Centaurs* (see Sec. 1.4). Observations made in 1988–89 showed that it exhibits cometary behavior and revealed

its coma by imaging. This fact prompted an action and led to a parallel, permanent cometary designation as 95P/Chiron. So far, there is one more Centaur with a dual status: (60558) Echeclus or 174P/Echeclus. A quite different object is comet 107P/Wilson-Harrington, discovered in 1949 but later considered as lost. In 1992, it was securely identified with Apollo asteroid (4015) 1979 VA, which was then named (4015) Wilson-Harrington.

1.3. Discovery Bias

As mentioned, comets develop their comae and tails in the innermost parts of their orbits, where they receive sufficient energy from sunlight. Thus, they brighten up considerably. In fact, the persistent part of a comet — the nucleus — is generally only several kilometers in extent. Until recently, such small objects had to come very close to the Earth in order to be discovered, and yet due to the comae and tails, comets are known since many centuries, and some were seen by the naked eye in prehistoric times. Occasionally, they are still able to dominate the night sky, while most comets observed in recent times remain very faint all the time.

It is thus clear that individual comets may be extremely different in brightness and ease of discovery. But it is also clear that the threshold brightness for a comet to be discovered has developed with time, as astronomy itself has developed, instrumentation has improved, and more people have become engaged in research including the search for comets. Nonetheless, there must always be a *discovery bias* as long as some comets remain too faint to be discovered. One of the important tasks when looking for statistical properties of the comet population is to limit this bias by an optimal definition of the observed sample.

While differences in brightness between comets may partly be due to differences in internal properties, they may also come from the different orbits. Since closeness to the Sun is so important for cometary activity, the *perihelion distance* (smallest distance from the Sun) is an essential orbital parameter. Statistically speaking, the larger the perihelion distance of a comet orbit, the less active the

comet will be, and the less likely its discovery. But, once again, this orbit-related bias has changed significantly over time.

Historically, the standard technique used for discovering comets has evolved dramatically. Centuries ago, it was a question of visual observations — by the naked eye or telescopic. Typical binoculars or telescopes used as comet finders had large fields of view and were thus very efficient for sky surveys, but their limiting magnitude was very bright so that only a minority of comets were within reach.

This situation gradually changed with the introduction of photographic techniques, using astrographs or Schmidt telescopes, for different purposes. Often the plates were aimed for star counts or probing extragalactic space, but they could also be scanned for moving objects like comets. Sometimes the aim of the observations was indeed to find comets or asteroids. The purpose would influence the sky coverage by the introduction of zones of avoidance or preference, depending on the nature of the targets. Galaxy counts would aim for high Galactic latitudes, while star counts would typically concentrate near the Galactic plane. Searches for asteroids would be made close to the ecliptic, avoiding the regions where the Galactic plane was crossed to avoid confusion with the rich stellar background.

Thus, an orbital bias could be introduced by an uneven sky coverage of the comet searches. This came on top of a fundamental source of bias due to the asymmetric distribution of observers between the northern and southern hemispheres of the Earth. The northern sky has been scanned much more efficiently than the southern sky — something that may, for instance, cause a bias in the distribution of arguments of perihelia.[4] Another obvious asymmetry arises from the fact that comets are most easily discovered on the night sky, favoring the opposition region. The discovery bias

[4]This is the angle measured along the plane of the comet orbit from a crossing point with the ecliptic (the ascending node) to the direction of perihelion. Together with the inclination, it determines the heliocentric ecliptic latitude of the comet, when it is at perihelion.

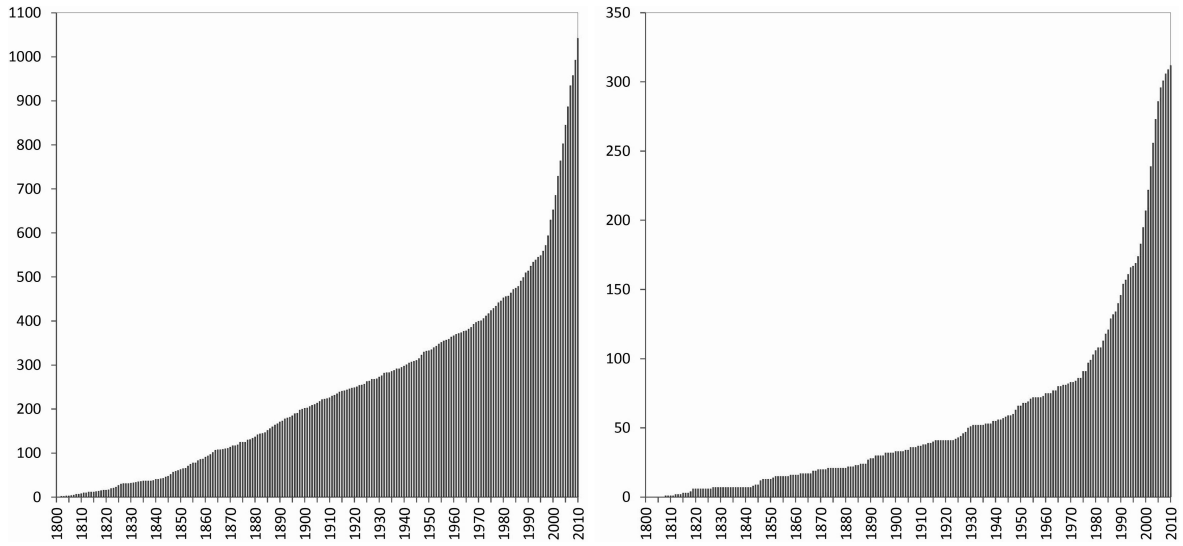


Fig. 1.1. Histogram representations of the cumulative number of discoveries during the years 1800–2010 referring to single-apparition comets (left panel) and numbered periodic comets (right panel). The list has been screened against comets with perihelion distances $q < 0.01$ AU or discovered by sun-staring space telescopes. The decrease of the discovery rate of periodic comets during the most recent years is due to the fact that, in general, two apparitions are needed before a comet is numbered. Hence, many of the recently discovered comets that will eventually be numbered are still counted as single-apparition comets. Courtesy T. Wiśniowski.

against comets passing perihelion on the side of the Sun opposite to the Earth is well known as the *Holetschek effect* since about a century ago.

Many of these biases have had a time dependence because of the changes in the ways comets are discovered. In particular, as illustrated by Fig. 1.1, the last few decades have seen a major rise in the rate of comet discoveries thanks to systematic search programs aimed at discovering Near Earth Objects (NEOs). This is obvious when glancing through the list of recent comet names, which is dominated by search projects like LINEAR, LONEOS, NEAT and Catalina or people associated with those. As a result, the total inventory of all ground-based comet discoveries is now dominated by those made recently by a small number of telescopes, located at a handful of sites with a certain geographical distribution and looking preferentially at low ecliptic latitudes. A special category is formed by the *sun-grazing comets* with perihelia close to the solar

surface, which hosts an enormous harvest from sun-staring space telescopes — primarily, SOHO.

Of course, it is very important to correct the orbital statistics of the discovered comets for the biases involved in the discoveries. This has always been an issue in cometary science. Anomalies found in the distribution of comet orbits may lead to suggestions of underlying processes like, for instance, an ongoing comet shower (see Sec. 5.3.1), or a planet orbiting in the Oort Cloud, and to judge the reality of those processes requires an analysis of discovery biases. Nowadays, it is fortunate that bias corrections can be straightforwardly applied using discovery simulators for a majority of comets using the logs or observing routines of the contributing surveys similar to what is done for NEOs.

In any case, the strong discovery bias in perihelion distance and its relation to the intrinsic brightness distribution of comets remains a major issue. Figure 1.2 shows how the perihelion distances of newly discovered comets have evolved during the past two centuries. The trends seen reflect the improved sensitivity and efficiency of comet discoveries. It is obvious that the perihelion distance distribution

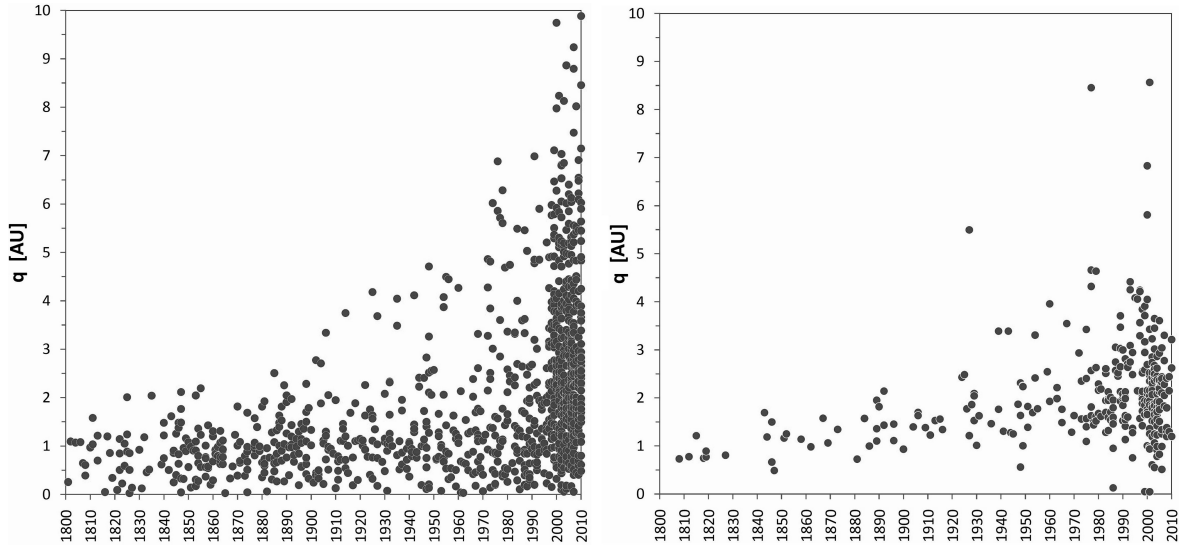


Fig. 1.2. Scatter plots of perihelion distance at discovery versus discovery year for comets discovered during the years 1800–2010: single-apparition comets (left panel) and numbered periodic comets (right panel) are plotted separately. The sample used is the same as in Fig. 1.1. A few comets with $q > 10$ AU are missing in each category. Courtesy T. Wiśniowski.

of the true comets has not undergone any major change in such a short time, so there have been effective limits in perihelion distance, beyond which comets escaped detection, or detections were limited by the brightness or activity of the comets. As time proceeded, these limits have moved outwards, and they are likely continuing to move.

A special case of discovery bias holds for the sungrazing comets. Before 1979, there were less than ten such comets on record throughout the history. These were extremely bright and overwhelming as a rule and frequently named "Great Comet" for the lack of only one or a few identifiable discoverers. Their brightness came from the fact that they passed perihelion at less than 0.01 AU from the center of the Sun. Only few such comets have been seen in the recent years, including the marginal case of comet C/2012 S1 (ISON). However, thousands of sun-grazing comets have been observed by a series of space observatories (SOLWIND, SMM, SOHO, Stereo), the most prolific of which was SOHO (SOlar and Heliospheric Observatory). These were seen passing perihelia or colliding with the Sun on coronographic images of the inner corona, but their faintness would have prohibited discovery by any other means. Most of them are tiny fragments of a common parent that also gave rise to the brightest sungrazers (Sec. 4.4.1). If included into the orbital statistics of comets, they completely skew the distributions of elements, and so they are excluded from the distributions presented here.

1.4. Comet Populations

Comets are often divided into separate populations. This of course concerns the observed comets, but we shall see that the observed comet populations tend to be associated with different, basically unobserved parent populations of objects with much larger perihelion distances. While the above-described categories like periodic comets and single-apparition comets are useful for book-keeping purposes, the populations are defined with an eye to the dynamical transfer routes that comets are expected to follow. One might ask if comets could be divided into different groups based on their morphologic

appearance or physical-chemical properties as well. However, this has proved remarkably difficult. One case in point will be introduced in Sec. 2.6.3.

As mentioned, an early classification was made into *short-period comets* and *long-period comets*. The limit was arbitrarily put at $P = 200$ years. This would in principle give the short-period comets a chance of being observed at more than one apparition (or perihelion passage), while the long-period comets would not stand such a chance. However, this division does not have any obvious dynamical significance. The orbital period of a comet is a very unstable quantity due to perturbations, primarily by Jupiter. Thus, from early times, there was a concept of *comet capture*, whereby Jupiter in particular would transfer comets from the long-period into the short-period class by decreasing the periods. Of course, the opposite can also occur, but since some short-period comets were known to have disappeared, their number might be kept more or less constant by a predominance of captures. Long-period comets would hence be the source of short-period comets.

We now know that reality is much more complex. Of particular importance is the recognition that short-period comets can be divided into groups that are relatively stable in dynamical terms. This builds on theoretical results within the *circular restricted three-body problem* (see Sec. 3.1), which were derived more than a century ago. There is an algebraic function of the semi-major axis a, the eccentricity e, and the inclination i of a comet orbit, called the *Tisserand parameter*, which is quasi-conserved under the perturbations by a particular planet, to which it is referred. For most comets, the perturbations by Jupiter dominate over those due to all the other planets, and hence we use the jovian Tisserand parameter

$$T_J = \frac{a_J}{a} + 2\sqrt{\frac{a}{a_J}(1 - e^2)}\cos i, \tag{1.1}$$

where a_J is the semi-major axis of Jupiter's orbit. As long as the perturbations by other planets are insignificant, T_J remains approximately constant over considerable time intervals, even though each

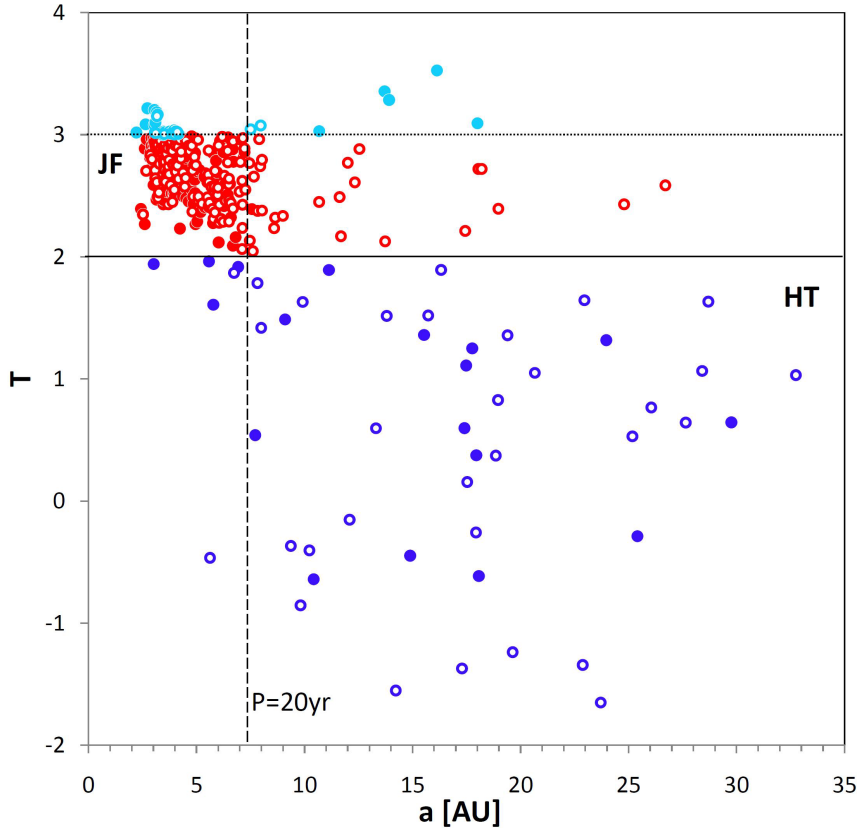


Fig. 1.3. Jovian Tisserand parameter versus semi-major axis of the short-period comets known in 2010, using their discovery orbits. Open circles denote single-apparition comets, and filled circles denote numbered periodic comets. The open circles are plotted on top of the filled circles in case of overlap. Blue symbols show Halley Type comets, red symbols show Jupiter Family comets, and sky-blue symbols are used for comets of Chiron or Encke type. The right-hand border of the diagram corresponds to an orbital period close to 200 years. Courtesy T. Wiśniowski.

of the elements entering into its definition may change dramatically as a result of close encounters with Jupiter.

Figure 1.3 shows how the comets with $P < 200$ years, known in 2010, are distributed with respect to a and T_J. By orbital evolution, these comets move essentially in the horizontal direction, so the diagram reveals that there are two, nearly disconnected main groups of short-period comets. The one with $2 < T_J < 3$ is in essence what we call the *Jupiter Family*, while the comets with $T_J < 2$ are usually called *Halley Type comets*. These two groups represent two distinct populations, which are not hermetically isolated but separated by a boundary ($T_J = 2$) that is difficult to cross. This boundary was introduced by Carusi and Valsecchi (1987) and is fundamental to the

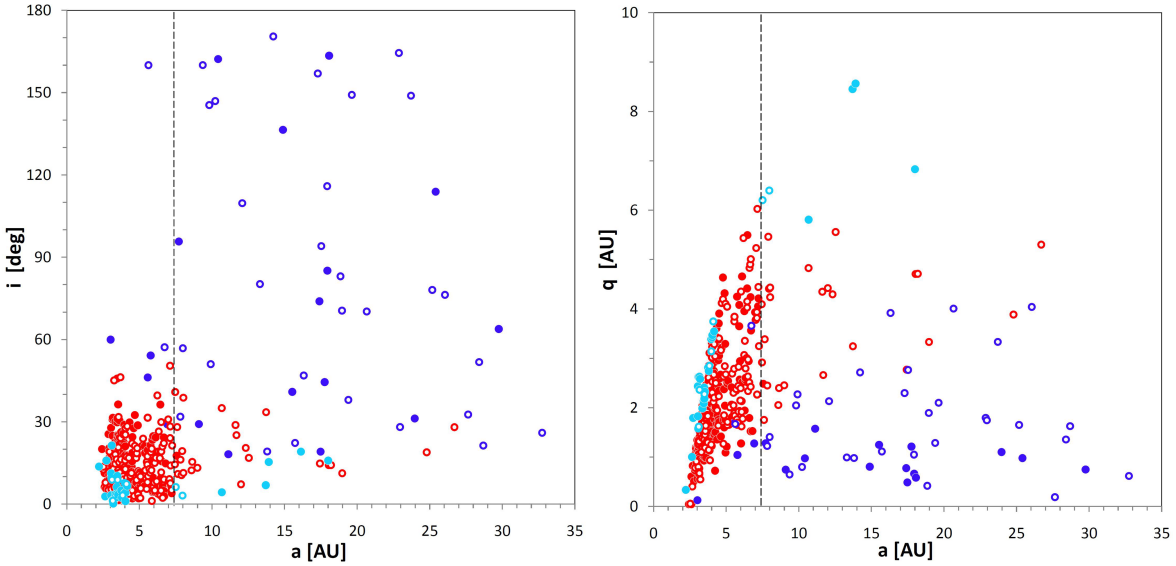
orbital taxonomy of comets introduced by Levison (1996) and further developed by Duncan *et al.* (2004). Here, all comets are considered irrespective of the orbital period, and those with $T_J > 2$ are called *ecliptic comets* due to their typically small inclinations, while those with $T_J < 2$ are called *nearly isotropic comets* to reflect the nearly isotropic distribution of their orbital poles.

In Fig. 1.3 we apply the taxonomy used by Duncan *et al.* (2004) by identifying as a special group the comets with $T_J > 3$. Those are of two different kinds. With $a > a_J$, they are referred to as Centaurs, of which there are many, but the objects in question have cometary designations due to their observed activity. They are also called *Chiron-type comets*. Those with $a < a_J$ and $T_J > 3$ are called *Encke-type comets* after their prototype, comet 2P/Encke, whose orbit is decoupled from Jupiter due to its small aphelion distance. This is quite remarkable, because in the current paradigm regarding the dynamical transfer of comets in the solar system (to be described in Sec. 3.6), comet orbits are either external to Jupiter's orbit in the case of remote reservoirs and the routes leading from these, or Jupiter-crossing in the case of observable comets. Orbits that fall entirely well inside that of Jupiter have to be seen as anomalous.

But nothing in our world is perfect. In particular, the term *Centaur* is used in a wider context for small bodies with orbits in the zone of the giant planets, but no exact definition has been agreed upon. The matter will be discussed later, but it is preferable to use the term Chiron-type comets rather than Centaurs for the comets in question. Moreover, the term Jupiter Family is usually thought to imply a dynamical control by Jupiter due to the possibility of close encounters. However, Fig. 1.3 shows a number of Chiron-type and Jupiter Family comets with similar semi-major axes, which straddle the dividing line at $T_J = 3$. The adherence to one group or the other is then a matter of orbital inclination and does not uniquely tell, if an object is under Jupiter's control or not. Finally, a recent category is formed by the so-called *Main Belt Comets*, whose orbits are typical of the asteroid main belt but which exhibit cometary activity. With the above definition, these are Encke-type comets but should really be treated as a distinct group.

A remarkable feature is that the two main groups (Jupiter Family and Halley Types) are almost disjoint also with regard to semi-major axis. Very few Halley Type comets have periods less than 20 years, while periods exceeding 20 years represent a minority in the Jupiter Family. A few decades ago the separation of the groups in orbital period was even more clearcut, and thus it was common to reserve the term short-period comets for those with $P < 20$ years and use *intermediate-period comets* for the range of periods between 20 and 200 years. This was almost equivalent to the distinction between Jupiter Family and Halley Type comets. However, as Fig. 1.3 demonstrates, the intermediate-period group nowadays has an important admixture of Jupiter Family and Chiron-type comets.

Figure 1.4(a) illustrates the major difference between ecliptic and nearly isotropic comets by means of the Jupiter Family and Halley Type comets. The former have inclinations that are almost always smaller than 30°, while the latter span the whole range up to 180° without any zone of avoidance. Among comets with periods



(a) (b)

Fig. 1.4. (a) Inclination versus semi-major axis, and (b) perihelion distance, versus semi-major axis for the comet apparitions plotted in Fig. 1.3. The symbol types and colors are the same. The vertical dashed lines indicate orbital period $P = 20$ years. One comet missing from the right panel is the Chiron-type comet 167P/CINEOS with $q = 11.8$ AU. Courtesy T. Wiśniowski.

$P < 20$ years, the few Halley Types have inclinations that are mostly close to the highest values of the Jupiter Family comets (30–60°), but one of them stands out by a strongly retrograde orbit. This is P/2006 R1 (Siding Spring).[5] At first sight, it looks like an extreme case of a Halley Type comet, while the others may rather be related to the Jupiter Family. However, the real nature of these objects cannot be judged without performing a comprehensive, dynamical investigation.

Regarding periods $P > 20$ years, the Jupiter Family and Chiron-type comets occupy the same range of low inclinations as some of the Halley Type comets. However, their orbits are obviously different because of the different values of T_J, and this difference is illustrated by Fig. 1.4(b). The perihelion distances of the Halley Types are mostly small, very rarely exceeding 3 AU, while the Jupiter Family comets show the opposite behavior. These follow the trend, seen among the smaller periods, for the perihelion distances to statistically increase with the orbital period — a natural consequence of the limited ranges of T_J and $\cos i$. It is now clear, why the number of Jupiter Family comets thins out for periods larger than 20 years. The large perihelion distances effectively counteract discovery of such comets.

Let us now consider the Encke-type comets. In Fig. 1.4(a), these mix perfectly with the Jupiter Family, and in Fig. 1.4(b) they appear to cling to the Jupiter Family on its left side. A better resolution of this orbital domain is provided by Fig. 1.5, which shows the (Q, q) plane of aphelion and perihelion distances with the line $Q = q_J$ indicated. This shows that the $T_J > 3$ comets are of three kinds. The Main Belt Comets are found at perihelion distances between 1.8 and 2.6 AU with aphelion distances less than 4.3 AU. They are well separated from the Jupiter Family, belonging to the orbital realm of the asteroid main belt. The only two comets that can indeed be called Encke-type are identified using their names: 2P/Encke (E) and 107P/Wilson-Harrington (W-H). These too are dynamically

[5] After 2010, two more comets have been discovered in rather similar orbits: 333P/LINEAR and P/2013 AL76 (Catalina).

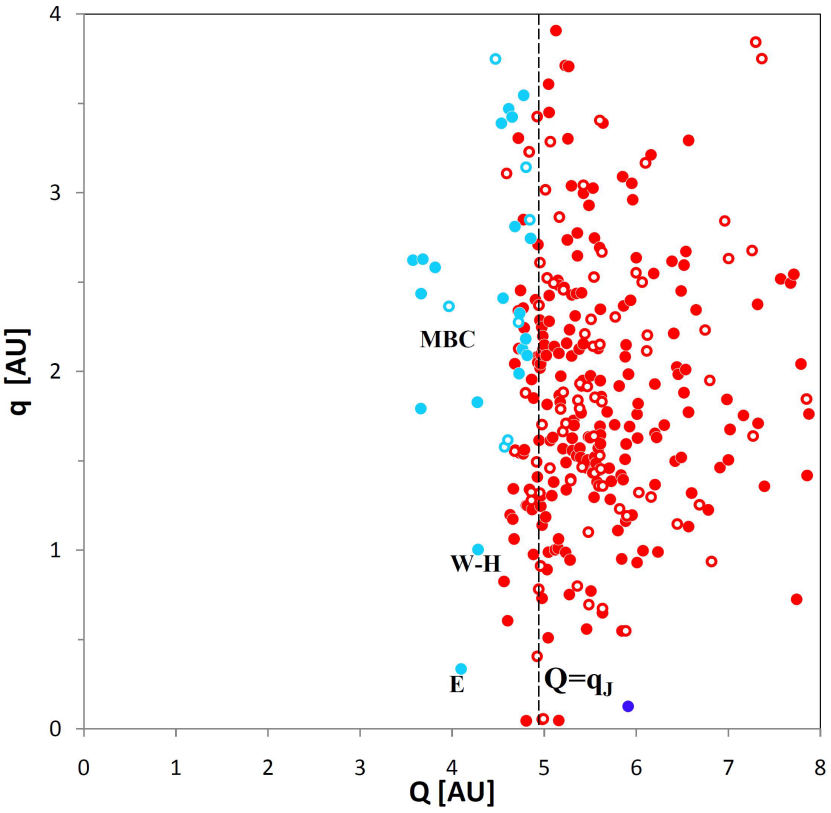


Fig. 1.5. Perihelion distance (q) versus aphelion distance (Q) for comets with $q < 4$ AU and $Q < 8$ AU. The meaning of the symbols is the same as in previous figures. The Encke-type comets with our definition ('E' and 'W-H'), and the Main Belt Comets ('MBC'), are identified (see the main text). Courtesy T. Wiśniowski.

decoupled from Jupiter due to their low aphelion distances. The rest of the comets are best counted with the Jupiter Family. In general, they mix with the low-Q border of the Jupiter Family, and at the highest q values (3.4–3.8 AU) there is a group of comets near the 3:2 mean motion resonance with Jupiter with a special dynamical history, which is considered a special subgroup of the Jupiter Family (see Sec. 3.2.2).

Finally, as will be detailed in Chaps. 3 and 5, the long-period comets form a more homogeneous group than the short-period ones. While the Jupiter Family and Halley Type short-period comets are considered to be derived from separate source populations in the outer solar system, no such subdivision can be made among the long-period comets. Their ultimate, distant source is considered as unique and common to all, namely, the *Oort Cloud*. This may also

be the main source of the Halley Type comets, while the Jupiter Family mainly stems from the *Scattered Disk*. We shall not deal with the source populations here, but these will be central topics of the discussions in Chaps. 5 and 6. One group of closely related long-period comets stands out as very important in terms of numbers but does not represent a separate population. This is the *Kreutz group* of sun-grazing comets, which are fragments of one and the same parent comet that has split due to the tidal force of the Sun (see Sec. 4.4.1).

1.5. Lessons from Space Missions

Chapter 2 will deal with the physical and chemical properties of comets as a background to the discussion of how they may have originated. However, this will focus on our knowledge about comets in general. It is obvious that this knowledge has an essential input from the close-up studies of comets that have been performed within different space missions. Let us thus first briefly review what has been learned or glimpsed from these projects concerning a small number of individual comet nuclei. These descriptions are not meant to be comprehensive, and more information will be given in later chapters.

1.5.1. *The Halley flybys in 1986*

Comet 1P/Halley was observed in times that were prehistoric in most countries, where comet science is pursued today. Its periodic nature was established through orbit determinations by Edmund Halley in 1705. Largely by setting this historic example of scientific progress through Newtonian mechanics, but also by the fact that it has often been a remarkably bright comet, 1P/Halley has become legendary. At its long awaited return to perihelion in 1986, it was therefore a natural choice as target of the first space missions to the near-nucleus environment of a comet.

In fact, there was a whole armada of such spacecraft. The most important was ESA's first interplanetary mission, called *Giotto* after the Italian, early Renaissance artist Giotto di Bondone, who may have been inspired by comet Halley when painting the *Adoration*

of the Magi, still to be seen in the Scrovegni chapel in Padova. Another pioneer achievement was made by the Japanese space institute ISAS, which launched two space probes called *Suisei* and *Sakigake* into the Halley environment. Last but not least, there were also two spacecraft of the USSR Academy of Science with French contributions, called *Vega 1 and Vega 2*, which first sent balloons into the Venus atmosphere and then were deflected toward Halley's comet by Venus' gravity (the name Vega comes from the Russian names of the targets, Venera and Galley).

Comet science was overthrown by the results of these encounters, mostly due to Giotto and the two Vegas. The most fundamental discovery was not unexpected, namely, that the comet has a solid nucleus. The most famous image of this nucleus, reproduced in Fig. 1.6, was taken by the Giotto HMC instrument (Halley Multicolour Camera). Even though the solid nucleus had been discussed since decades, this result was of paramount importance, since the hypothetical concept of the comet nucleus that Fred Whipple (1950) had introduced was now confirmed, and all lingering doubts could be disposed of.

Fig. 1.6. Composite image of the 1P/Halley nucleus acquired by the Halley Multicolour Camera in March 1986. Reproduced with permission from H. U. Keller.

More surprising was the size of the nucleus. Ever since the recovery of the comet in 1982, observations had been performed, which estimated the *photometric cross-section* of the nucleus. This is a measure of the product of the albedo and the geometric cross-section — the latter measuring the square of the average radius. In 1986, comet albedos were not known and models often assumed very high values, since the scientists' minds were fixed on a slightly contaminated snowball. With such an assumption, the radius of the Halley nucleus would be much smaller than the images revealed, and the result was an unexpectedly low albedo. In fact, this distinguished the only explored comet nucleus as one of the darkest solar system small bodies ever observed.

The large size of the nucleus had one more unforeseen consequence. The water production rate of comet Halley would have been much higher than observed, if its nucleus had been the expected snowball-like object. In fact, the flow of water vapor due to sublimation must be quenched over the whole surface or at least the majority of it. This could be consistent with the HMC image, which shows jet-like features of dust emerging from what seems like local, active spots. However, this interpretation has not been confirmed and may be an oversimplification. What is certain is that the surface has to be enriched in refractory substances, or the whole nucleus is made of a dusty matrix with included ice rather than the other way round. This is in agreement with the high surface temperatures observed on the Halley nucleus, reaching a level around 350 K.

In fact, thanks to the dust detection systems on board Giotto and Vega, a further important conclusion could be drawn. The grain size distribution turned out to fall off less steeply toward macroscopic chunks than the earlier models had predicted. As a consequence, the total mass of the ejected dust was inferred to be at least as large as that of the outflowing gas. Fragmentation of the larger grains into smaller pieces on the way out through the coma was also observed. Overall, the "Halley armada" showed comet nuclei — by inference from the single object under study — to be bigger, darker, more dusty, and possibly more porous and less coherent than earlier believed.

The elemental composition of the comet dust was studied by means of mass spectrometry of single, small grains. These turned out to be of different kinds: one dominated by silicates, a second dominated by organic molecules rich in carbon, hydrogen, oxygen and nitrogen (the so-called CHON grains), and a third being a mixture of the two. Combining this with what could be inferred about the gas composition led to a picture of the overall elemental composition of the material ejected from the 1P/Halley nucleus, which is illustrated in Fig. 1.7. Comparing this to the composition of the most primitive (i.e., solar-like) meteorites — the CI chondrites — shows the comet to be even more primitive.

This finding highlighted a property of comets that was not unexpected though extremely important. They appear to be unaltered, comprehensive collections of elements from solar nebula regions,

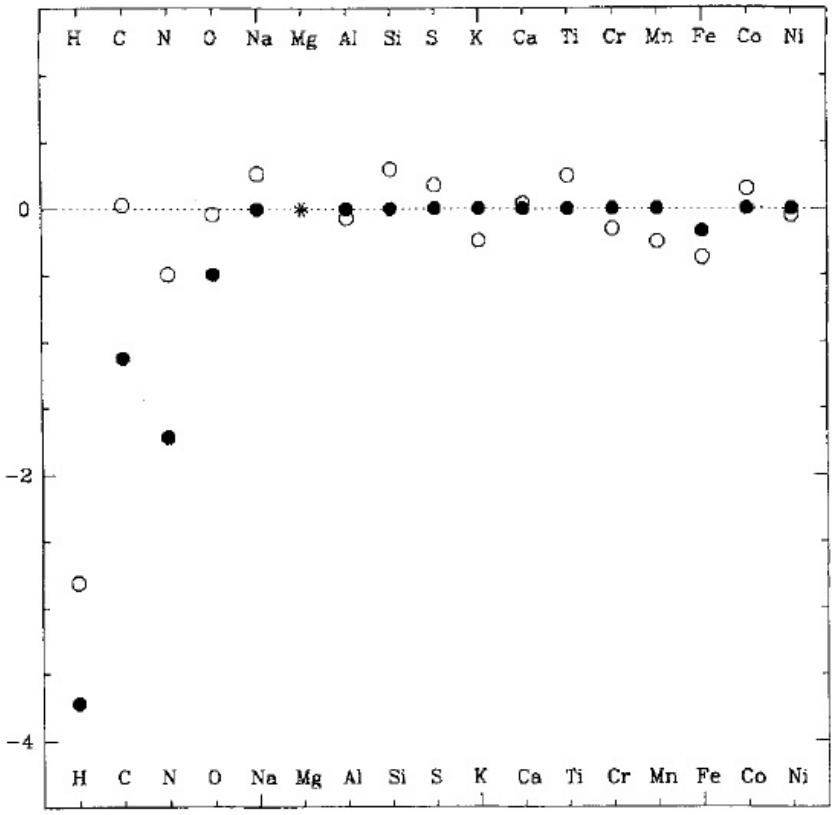


Fig. 1.7. For a suite of cosmically abundant elements, the ratios between abundances in CI chondrites and the Sun (filled circles), and in comet 1P/Halley and the Sun (open circles), are shown on a log scale. The reference element for the abundances is magnesium (Mg). From Festou, M. C. *et al.*, *Astron. Astrophys. Rev.* **5**, 37–163 (1993), © Springer-Verlag. With permission of Springer.

where all condensible material was indeed condensed, i.e., very cold regions. Further support for this came from the measured isotope ratios in grain material. Some isotope ratios in comet Halley (e.g., $^{12}C/^{13}C$ and $^{32}S/^{34}S$) are very close to the solar system average, presented by the Sun's photosphere and the most primitive meteorites. At the same time, large variations of isotope ratios in individual, microscopic grains exist, showing that pre-solar units may have been preserved.

1.5.2. *The flybys of 19P/Borrelly and 81P/Wild 2*

Comet 19P/Borrelly was discovered by French astronomer Alphonse Borrelly more than a century ago, and with a period of a little less than 7 years it has been observed at most returns since then. It is an average Jupiter Family comet that has not shown any remarkable behavior. However, it presented an attractive opportunity for NASA to steer its *Deep Space 1* technology test mission to a comet target after it had accomplished its prime scientific goal — the flyby of asteroid (9969) Braille. The flyby of 19P/Borrelly took place on 21 September 2001, when the comet was close to its perihelion at about 1.36 AU from the Sun.

The DS1 spacecraft carried much less instrumentation for comet exploration than Giotto and the Vegas did, but because this was the first visit to a comet after 1986 and the imaging conditions were much more benign than they had been for comet Halley,[6] the images taken of 19P/Borrelly had great significance. One of them is shown in Fig. 1.8. This nucleus is smaller than that of 1P/Halley, measuring about 8×3 km. The two nuclei are similar in their elongated shape, and the Borrelly nucleus was likened to a footprint. Just like the Halley nucleus, it was found to be extremely dark — in fact, it beat the record of darkness for surfaces of solar system objects with a mean Bond albedo of 0.01. Another similarity is that the global

[6]Due to the strongly retrograde orbit of 1P/Halley, the spacecraft encountered the comet with a speed of about 70 km/s, while the low-inclination orbit of 19P/Borrelly allowed the DS1 encounter to happen at only 16.5 km/s.

Fig. 1.8. Image of the 19P/Borrelly nucleus acquired from the Deep Space 1 spacecraft on 21 September 2001. Reproduced with permission from L. Soderblom. Credit: NASA/JPL-Caltech/US Geological Survey.

outgassing rate observed during the DS1 encounter was much smaller than expected for an icy surface of the same size.

The morphologic appearance of the surface is seen to be variegated. Several different units have been identified, and the local albedo varies by a factor of almost four. While the overall topography is rough, some local units appear relatively smooth. No impact structures have been identified. A possible indication of a binary nature from the overall shape remains speculative, and there is no sign of the "sole" and "heel" of the footprint having different morphologic characters.

Near-infrared spectroscopy of the 19P/Borrelly nucleus was also performed by DS1. The most noteworthy conclusion was that no trace of surface ice was seen in these spectra, although this had been an easy matter, if the surface had been rich in ice.

The next comet mission performed by NASA would be much more scientifically ambitious. This was the *Stardust* mission, launched in February 1999. Stardust was aimed to collect coma grains in comet 81P/Wild 2 in silica aerogel during a close approach to the

comet nucleus on 2 January 2004, and bring them back to Earth for laboratory analysis. All went well, and the Stardust harvest of more than 10 000 comet particles represents one of the most important sources of information available on the formation conditions of comets.

Comet 81P/Wild 2 was discovered in 1978 by Swiss astronomer Paul Wild. It is a Jupiter Family comet with an orbital period of 6.4 years, but as such it is very young, and this is its distinguishing property. In September 1974, resulting from a close encounter with Jupiter, it was captured into the current orbit from a larger perihelion distance and much longer period. This may mean that it is less physically evolved and therefore relatively fresh, compared to most Jupiter Family comets — an attractive property for a space mission target.

Images of the Wild 2 nucleus were acquired by the Stardust navigation camera, and one of them is shown in Fig. 1.9. Contrary to what is often stated, its shape is not markedly oblate, but nor is it as prolate as the Halley and Borrelly nuclei. Its dimensions are

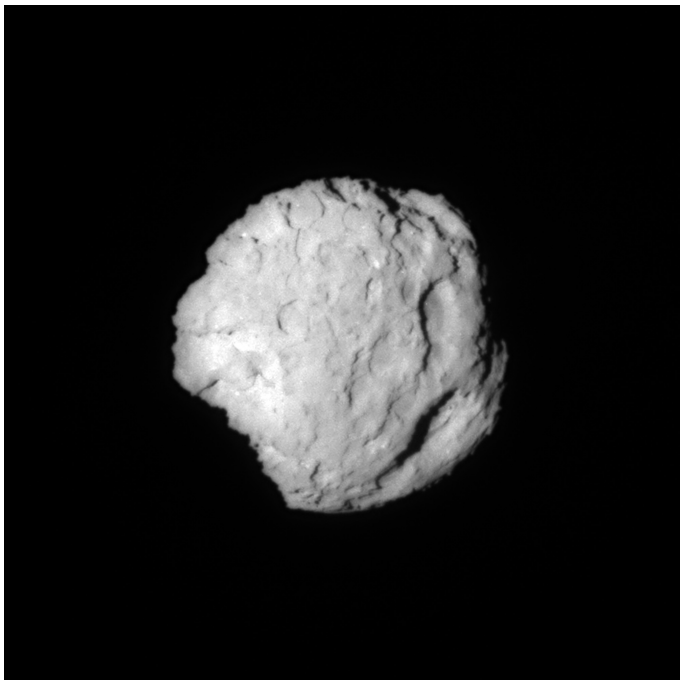


Fig. 1.9. Image of the 81P/Wild 2 nucleus acquired by the Stardust camera. Reproduced with permission from D. E. Brownlee. Credit: NASA/JPL-Caltech/University of Washington.

$5.5 \times 4.0 \times 3.3$ km. Similar to Halley and Borrelly, it has a low albedo estimated to be close to 0.03.

The surface features on the Wild 2 nucleus include elements that were not seen on the preceding mission targets. These may indeed provide insights into the evolution of the nuclei. The most characteristic features are circular depressions. These are of two kinds: pit-halo features and flat-floor features, ranging in size from 0.5 to almost 2 km. According to Brownlee *et al.* (2004), they are most likely impact structures. Their morphologies do not coincide with those of impact craters on rocky solar system objects, but the nature of the target and its material are also very different. If the inference is correct, one may conclude that the lack of impact features on the Borrelly nucleus is due to their destruction on a time scale similar to the typical residence time of Jupiter Family comets in orbits like those of Borrelly and Wild 2.

However, the most remarkable results from the Stardust mission concern the composition of the sampled coma grains. These will be discussed in later chapters and are only briefly summarized here. An unexpected result is a strong heterogeneity of the olivine and pyroxene compositions between individual grains, which indicates very different formation conditions. Thus, on a micrometer scale, comet Wild 2 appears to host particles from all over the solar nebula, which necessitates efficient radial mixing from the innermost parts into the trans-planetary region, where comets were likely formed.

In particular, high-temperature minerals like forsterite and enstatite are present along with much less refractory components like iron-magnesium sulfides and organics. Even a CAI-like[7] composition has been observed for one grain. Another remarkable finding is that the compositional evidence seems to exclude acqueous alteration in comet Wild 2, contrary to the case of most carbonaceous chondrite meteorites — see Sec. 8.1.2. Finally, even though the isotope ratios of the most abundant elements show a wide scatter between individual

[7]CAI stands for Calcium–Aluminum-rich Inclusions. These occur in many chondritic meteorites and consist of minerals with extremely high condensation temperatures, often involving calcium and aluminum.

grains, the absence of extreme anomalies indicates that presolar material is rare in comet Wild 2.

1.5.3. *The study of comet 9P/Tempel 1*

The next comet to be investigated by spacecraft flybys was 9P/Tempel 1. It was discovered 150 years ago by the German astronomer Wilhelm Tempel. After the discovery apparition in 1867 it was observed at its next returns in 1873 and 1879 but was then lost for nearly a century. During that time it was sometimes deemed to have disappeared, never to be found again, but the history of comet science knows several examples of a memorable fact: comets should not too easily be declared dead! Comet Tempel 1 was photographed again in 1967 and definitely recovered in 1972 after painstaking calculations by British astronomer Brian G. Marsden and has never been missed since then.

In this case, the reason for the long interruption of the observations has to do with the orbital evolution of the comet. It repeatedly undergoes moderately close approaches to Jupiter in connection with a libration around the 2:1 mean motion resonance with the planet (see Sec. 3.5). This means that the comet spends intervals of about a century on either side of this resonance, and the approaches occur during the transits. When Tempel 1 is outside the resonance (semi-major axis larger than one half that of Jupiter), the perihelion distance is close to 2 AU, and this was the case from the 1880s to the 1950s. In recent times, the comet has had $q \simeq 1.5$ AU like it had, when it was discovered. This librating motion can be traced with reasonable confidence back to a close encounter with Jupiter in 1703 (Yeomans *et al.* 2005), before which the perihelion distance was apparently much larger.

Comet Tempel 1 was chosen as the target of NASA's *Deep Impact* mission. This carried a spacecraft designed to perform close-up observations of the comet nucleus and an impactor, which was disconnected from the main spacecraft a little before the encounter and steered onto collision course with the nucleus. The aim was to excavate nuclear material from some depth to observe its composition

and thereby to better understand the relationship between the well-observed coma abundances in many comets and the make-up of the nuclear material from which the coma species originate. The operations were basically flawless. On 4 July 2005, this man-made little cosmic impact took place, delivering a kinetic energy of 19 GJ to the selected place on the comet nucleus.

The Deep Impact (DI) spacecraft would not return to comet Tempel 1. However, on 14 February 2011, this comet was revisited by the Stardust spacecraft, which had been steered to this goal using an Earth swing-by in 2009 and several other trajectory correction maneuvers in what was called the *Stardust-NExT* mission (Veverka *et al.* 2013). The Tempel 1 nucleus is so far the only one that has been visited by spacecraft more than once.

Figure 1.10 shows two pictures of the Tempel 1 nucleus — one from each of the visits. The combination of the two sets of images was important to determine the dimensions of the nucleus as 7.9×4.2 km, since the spin period is relatively long (40.7 hours). Once more, like for the preceding targets, this object was seen to be of low albedo and very large for its observed gas production rate.

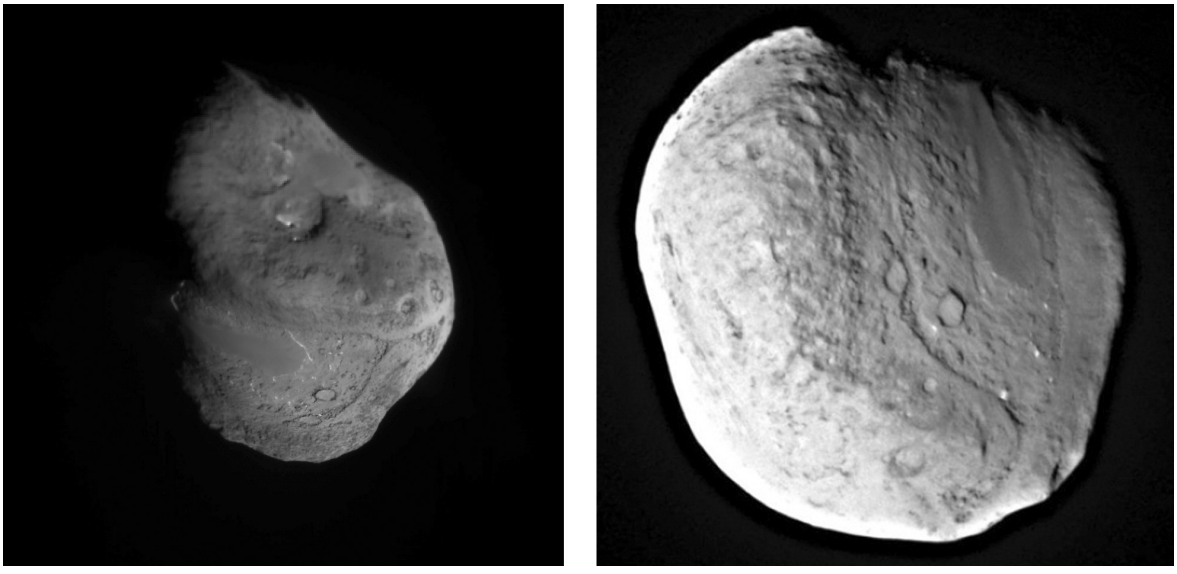


Fig. 1.10. Left panel: Image of the nucleus of comet 9P/Tempel 1, acquired by the Deep Impact spacecraft. Reproduced with permission from M. F. A'Hearn. Credit: NASA/JPL/University of Maryland. Right panel: Image of the same nucleus, acquired more than five years later by the Stardust NExT spacecraft. Credit: NASA/JPL/Cornell University.

It is natural to compare the surface morphologies between the two well-imaged comets, Wild 2 and Tempel 1. Concerning the circular features that are so prominent on Wild 2, such features are also seen on Tempel 1. In this case they appear more subdued and eroded, but an origin as impact craters cannot be excluded. One may wonder, if the surface erosion that has occurred on Tempel 1 since 1703 is on its way to wipe out the memory of a past collisional evolution, which is still prominent on Wild 2.

Let us not forget that round depressions, or pits, can have an endogeneous origin by local, jet activity as well as being impact craters. Time scales are of relevance for this discussion. Can the pits observed on the two comets be formed by collisions with interplanetary boulders in the time available? This time may be limited by both the dynamical residence time in a collisional environment and the surface erosion of comets with small perihelion distances. For Tempel 1, Belton *et al.* (2013) found the impact origin to be unlikely by comparing the number of pits with an expected dynamical lifetime of ∼30 000 years or less for Jupiter Family comets. However, the dynamics in question allows for the possibility of extended stays in orbits with large perihelion distance and minimal erosion, so the issue remains open.

Even more attention has been paid to a morphological feature that was discovered on the Deep Impact images of Tempel 1, while evidence could *a posteriori* be traced on the Borrelly and Wild 2 images too. This is referred to as *layering*. Two different manifestations were identified (Thomas *et al.* 2007). One is a set of linear outcrops running more or less in parallel across the nucleus over a length of 3.5 km, forming bands of width ∼10–100 m. These are interpreted as surface manifestations of layers penetrating deep into the interior of the nucleus. The other is a region of flat surface with scarps of ∼1–10 m height. These scarps are interpreted as erosional features, formed when surface mass loss proceeds along thin layers rather than perpendicular to them.

The near-IR spectroscopic studies made by Deep Impact have brought important information about the water production activity of comet Tempel 1. Temperature maps of the nucleus have been

constructed, showing that the surface temperature is far above the level expected for sublimating H_2O ice (like earlier observed for the 1P/Halley nucleus). This means that the flux of water molecules leaving the nucleus in the near perihelion part of the orbit must be maintained essentially by subsurface sublimation. A spectral search for icy patches on the Tempel 1 nucleus by Sunshine *et al.* (2006) revealed only very small areas, far too small to sustain the observed H_2O outgassing. This confirmed the DS1 results from 19P/Borrelly concerning the near absence of surface ice. Thus, active regions on comet nuclei, as often used in theoretical models, do not appear to be identifiable with icy areas of considerable extent.

The above-mentioned temperature maps have been derived using the short-wavelength part of the thermal emission from the Tempel 1 nucleus as captured by the DI imaging spectrometer together with modeling of the physical processes behind this emission. A good fit to the observational data may then imply not only a charting of the surface temperature but also a determination of the thermal inertia and surface roughness, if the observing geometry is favorable (Davidsson *et al.* 2015). The best effort in doing this for Tempel 1 (Davidsson *et al.* 2013) led to the conclusion that significant variations of the thermal inertia occur across the nucleus surface — some regions being measurably rough with extremely low inertia (less than 50 MKS units) and others reaching 3 to 4 times larger inertia values depending on roughness (though still very low). A low thermal inertia essentially means a low conductivity, which in turn implies a loose, fine-grained material with little contact between the grains.

The impact experiment brought several interesting results. Volatiles like H_2O and CO_2 were released in large amounts in the earliest ejecta, showing that they exist close to the surface. Organics were also abundant in the ejecta. The large volume of very fine particles that was ejected shows that these grains are an integral part of the nuclear material — either as individual units or as building blocks of loose aggregates. Moreover, this fine-grained material must exist at considerable depth as well as at the very surface. Finally, the analysis of the ballistic motion of the ejecta revealed that the

material had almost vanishing strength, and that the bulk density of the nucleus was very low, estimated at $\sim$400 kg/m^3. While of great importance, both the last mentioned results essentially confirmed already published findings (see Sec. 2.5.3).

1.5.4. *Observations of comet 103P/Hartley 2*

Comet 103P/Hartley 2 was discovered by English astronomer Malcolm Hartley in March 1986. He was using the Uppsala Schmidt telescope at Siding Spring Observatory in Australia within a program forming part of the US-based Catalina search for Near Earth Objects. Comet Hartley 2 is indeed such an object, whose perihelion distance at the time was 0.95 AU. Soon before its discovery, it had completed the third perihelion passage in such an orbit, but in 1971 it underwent a close encounter with Jupiter, which brought it closer to the Sun from a preceding perihelion distance of about 1.5 AU. Before another encounter in 1947, the perihelion distance was more than 2 AU (Carusi *et al.* 1995).

The main spacecraft of the Deep Impact mission had enough instrumentation left in good shape after the passage of comet Wild 2 that it was worth pushing it toward a second cometary target. Thus an extended investigation was approved under the name *EPOXI* (first called Dixi), and comet 103P/Hartley 2 was the target within reach. Among the cometary space mission targets thus far, it was probably the least explored, but its relative freshness from a dynamical point of view was of course a positive aspect. The encounter took place on 4 November 2010, one week after perihelion, and the main results came from the imaging and spectroscopic studies.

A picture of the Hartley 2 nucleus is shown in Fig. 1.11. This nucleus was known in advance to be quite small, and a good size determination using the Spitzer Space Telescope had been published (Lisse *et al.* 2009). The EPOXI imagery showed a bi-lobed shape of dimensions 2.33×0.69 km. The albedo is only 0.04 in spite of its fresh appearance. Ejection of material is prominent from the Sun-facing, minor lobe.

When comparing the water production rate of comet Hartley 2, as measured during the 2010 apparition, with the size of the nucleus, a

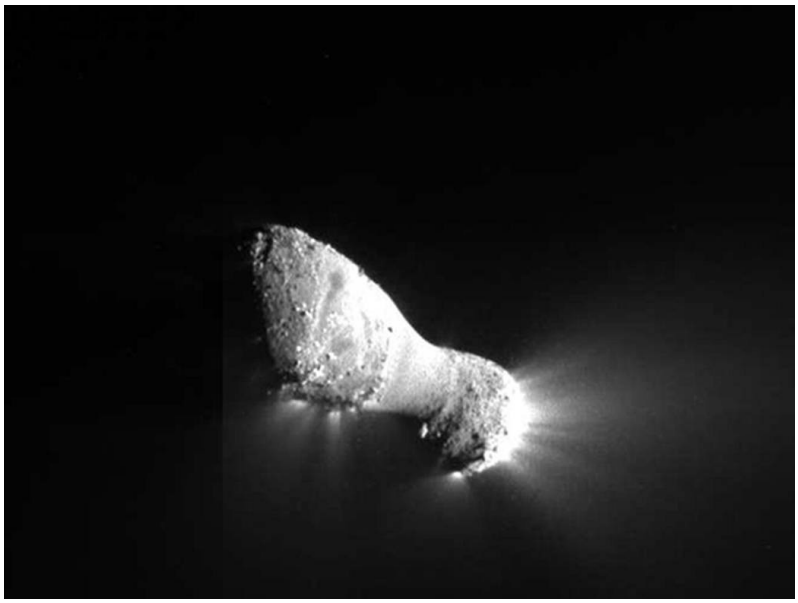


Fig. 1.11. Image of the nucleus of comet 103P/Hartley 2, acquired by the Deep Impact spacecraft within the EPOXI mission. Reproduced with permission from M. F. A'Hearn. Credit: NASA/JPL/University of Maryland.

remarkable feature stands out. This comet is hyperactive, producing significantly more H_2O than an equivalent sphere of pure H_2O ice at the same distance from the Sun. In this regard, it is quite different from all the other cometary spacecraft targets. The reason for the hyperactivity was revealed by the EPOXI observations. The prominent activity emanating from the small lobe is driven by sublimation of CO_2, which consequently is abundant in the near-surface material, at least in this part of the nucleus. This causes the release of large grains ("chunks") of nearly pure H_2O ice, which subsequently sublimate — thereby contributing a large part of the observed H_2O production rate (A'Hearn *et al.* 2011).

Hence, it is obvious that the outgassing activity of the Hartley 2 nucleus is not primarily indicated by the water production rate. Water was found to sublimate from the nucleus too, but mainly from the waist region and probably resulting from icy grains that were released from elsewhere and fell back to this area. Some degree of chemical heterogeneity, apparently between the two lobes, has been suggested with a variation of the CO_2/H_2O ratio by a factor two. The large abundance of CO_2 is remarkable, especially relative to CO. The latter molecule is less than 1/60 as abundant as the

former in comet Hartley 2, indicating that the material is much more oxidized than what has been observed in other comets (A'Hearn *et al.* 2012).

The satellite-based exploration of comet Hartley 2 during its 2010 apparition led to another remarkable result. This was the discovery, using the Herschel Space Observatory, of an ocean-like deuterium/hydrogen ratio in the water molecules of the inner coma. The submillimeter line of two isotopic variants of H_2O (HDO and $H_2^{18}O$) were observed simultaneously, and a value of 500 ± 50 was assumed for the $H_2^{16}O/H_2^{18}O$ ratio. From this, the D/H ratio was found to be consistent with the Vienna Standard Mean Ocean Water (VSMOW) value (Hartogh *et al.* 2011). Such ratios had been observed in several other comets, starting with 1P/Halley in 1986, and these had shown a more or less common behavior with an average D/H ratio close to a factor two larger than VSMOW. Since all the earlier observed comets belonged to the groups associated with the Oort Cloud, while Hartley 2 may rather be referred to the Scattered Disk (see Sec. 1.4), the difference in D/H ratio might reflect a difference in formation conditions at different distances from the Sun — a fundamental issue when discussing comet origins. We shall return to this in Sec. 8.1.3.

1.5.5. *The scrutiny of comet 67P/ Churyumov-Gerasimenko*

The ESA Rosetta mission to comet 67P/Churyumov-Gerasimenko represents a quantum leap in cometary exploration. Some of its results will be referred to in the following chapters, and here we only give a brief introduction. Comet 67P was discovered on photographic plates in 1969 by Soviet astronomers Klim I. Churyumov and Svetlana I. Gerasimenko, who were active at Kiev University. This is in most respects an average Jupiter Family comet except for the fact that it has been scrutinized by the Rosetta instruments, yielding information far beyond what has been obtained for any other such comet. Its perihelion distance at discovery was 1.29 AU and is now slightly smaller.

A close encounter with Jupiter in 1959 had transformed the orbit from a preceding perihelion distance of 2.76 AU. In this sense, the comet reminds us a bit of 81P/Wild 2. The latter case was more extreme, but comet 67P may also be seen as a relative newcomer in its present orbit. In view of the official aim of the Rosetta mission — to read the early history of the solar system from its imprint on the comet nucleus like the Egyptian hieroglyphs were interpreted due to the trilingual inscriptions on the Rosetta stone — this kind of freshness is definitely an attractive property. In this sense, comet 67P is an improvement over the original Rosetta target,[8] comet 46P/Wirtanen, whose recent dynamical history is more quiescent.

The mission scenario in itself is quite remarkable. After launch in March 2004 the spacecraft had to travel for ten years through interplanetary space, using three gravity assists at close encounters with the Earth and one with Mars. Thus, its heliocentric orbit was made similar to the comet orbit, and a very slow approach was performed relatively far from the Sun in 2014, after which a rendezvous followed, lasting about two years. Science data were obtained at the planetary encounters and especially at two main belt asteroid flybys: with (2867) Šteins in September 2008 and (21) Lutetia in July 2010.

The rendezvous with comet 67P started with arrival in August 2014, when the comet had passed its aphelion and was slowly approaching the Sun. The heliocentric distance was then 3.7 AU. Orbit insertion was finished within a few weeks. In November 2014, at 3.0 AU from the Sun, a lander called *Philae* was dropped onto the nucleus surface. The comet was then still of low activity, and the gradual increase until perihelion passage in August 2015 and decline thereafter were monitored from the orbiter with multiple instruments. The scientific use of Philae, however, was severely hampered by an unfortunate failure at touchdown, which made it

[8] The reason for the change of target comet was an Ariane 5 launch failure shortly before the planned Rosetta launch, which was thus delayed by about one year.

bounce a few times while travelling a long distance across the nucleus and come to rest without radio contact with the orbiter. For almost the whole remaining part of the mission, its exact location remained unknown. The mission ended spectacularly on 30 September 2016, as the orbiter was steered into a hard landing on the nucleus in a controlled manner.

One of the most remarkable properties of this nucleus is its binary nature, which is clearly displayed in Fig. 1.12 and will be a topic of discussion in Sec. 7.3.1. The most commonly used analogue from real life experience is a rubber duck — thus, the small and large lobes are often called the head and the body, respectively, while the connecting region is called the neck. Moreover, an intricate and varied surface morphology is seen to characterize the whole object. Some of these features, including temporal and episodic changes, will be presented

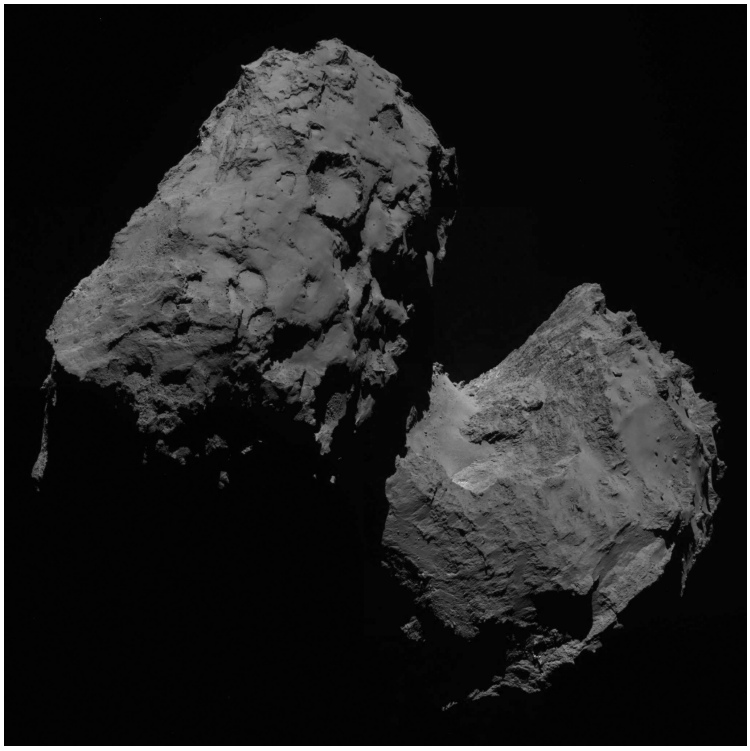


Fig. 1.12. Image of the nucleus of comet 67P/Churyumov-Gerasimenko, acquired by the OSIRIS camera onboard the ESA Rosetta spacecraft on 6 August, 2014. Credit: ESA/Rosetta/MPS for OSIRIS Team MPS/UPD/LAM/IAA/SSO/INTA/UPM/DASP/IDA.

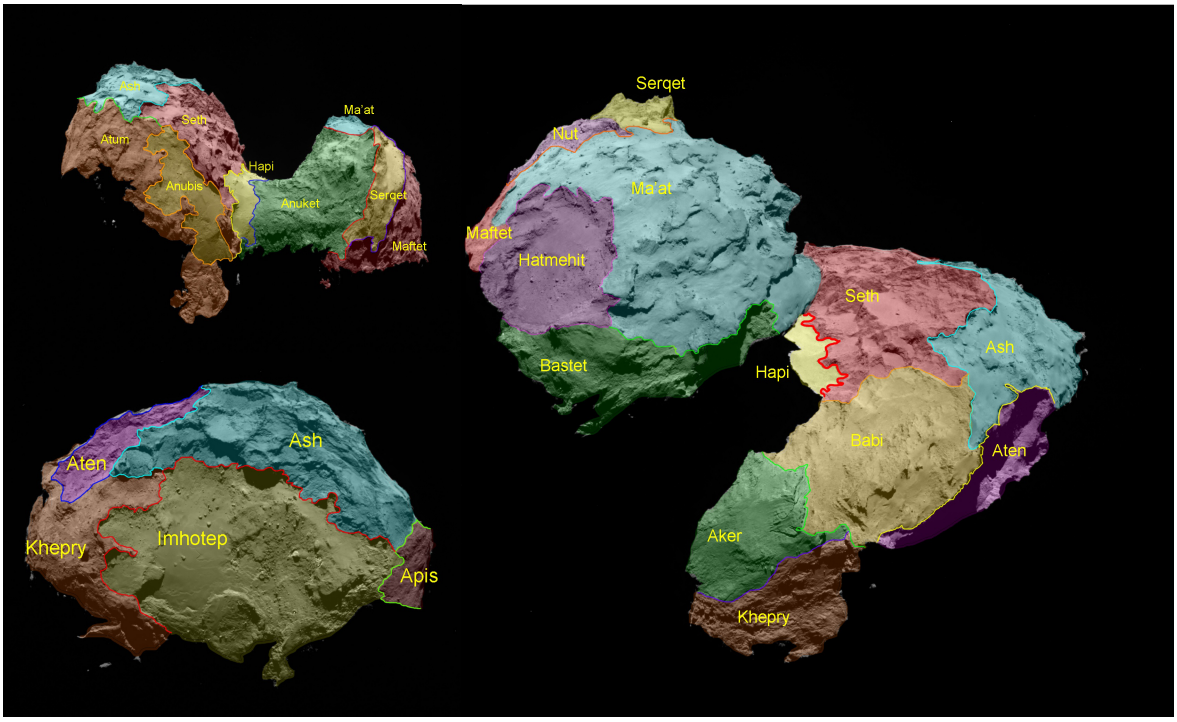


Fig. 1.13. Regional definitions for the 67P nucleus, corresponding to large-scale morphologic provinces. From Thomas, N. *et al.*, *Science* **347**, aaa0440 (2015). Reprinted with permission from AAAS.

in Chaps. 2 and 4. To prepare for those discussions, Fig. 1.13 shows the morphologic provinces defined by Thomas *et al.* (2015a), which are used to describe the different geologic formations and localize the individual features. This picture comes from the initial exploration by the OSIRIS cameras and thus shows only the part of the nucleus that was sunlit at that time.

This page intentionally left blank

Chapter 2

Physical and Chemical Properties

The first topic to discuss concerns the observational evidence regarding comet nuclei. Physical modeling is of importance to elucidate some of the properties and determine the parameters in question. However, the observations are of prime importance and will be dealt with in the first place. These are often difficult to perform or vulnerable to biases and systematic errors, so the results have to be discussed with a critical mind. It is natural for this to be reflected in the following descriptions, which sometimes tend to look a bit pessimistic. However, when we deal with the foundations of the ambitious and comprehensive theories on the role of comets in solar system formation and evolution, it is of essence to distinguish what we actually know from what appears to be favored by indirect evidence, and what we simply have to guess.

2.1. Size Distribution

One of the most important statistical properties of any comet population is the size distribution of the nuclei. This was initially shaped by the processes whereby the nuclei were formed, and it may later have been changed by the physical evolution of these nuclei, thus providing a record of these processes. Clearly, if we knew the size distributions of all kinds of comets with great precision, we would have an invaluable tool to unravel and understand the origin and evolution of comets.

Unfortunately, this is not the case. There are several reasons and the most important is that the detailed observations, from which the sizes and shapes of comet nuclei can be inferred, are very demanding. Hence, these are available only for a minority of comets. In particular, from close-up images of comet nuclei obtained using space missions (Sec. 1.5), the sizes and shapes are reasonably well known without ambiguities, but there are only a handful of comets that have so far been visited this way. This leaves remote sensing observations as the only way to deduce the sizes of all the other nuclei.

2.1.1. *Visual nuclear magnitudes*

Since the size distributions essentially rely on such remote observations, let us now see how these are performed. First of all, given the size range involved, it is hopeless to wait for comets to pass, by chance, close enough for their nuclei to be resolved by Earth-based imaging. We thus have to deal with unresolved objects. The usual way to infer the sizes of such objects in the solar system is by means of photometry at visual wavelengths. We need to observe the *nuclear magnitudes* of comets.

Specifically, the input is photometric data referring to the nucleus, obtained in a standard system and thus expressible as apparent visual magnitudes (m_V). Treating these magnitudes as truly nuclear, they can be used to infer an *absolute magnitude* H_N from the formula

$$m_V = H_N + 5 \log \Delta + 5 \log r + \delta m, \tag{2.1}$$

where Δ is the distance from the Earth and r is the distance from the Sun, both expressed in AU, and δm is the phase correction of the magnitude. Its dependence on the phase angle ϕ is sometimes called the *phase function*. Due to measurement errors, inaccuracies in the assumed phase function, and the fact that the nucleus may spin and have an irregular shape, the values of H_N obtained at different times for the same comet may differ, and thus it is preferable to use an average over many observations.

Lacking information about the actual shape, this average can then be assumed to correspond to an equivalent sphere[1] with radius R_N. The relation between R_N and H_N can be expressed as

$$\log\left(p_V \pi R_N^2\right) = 16.85 + 0.4(m_\odot - H_N), \tag{2.2}$$

where R_N is expressed in km. Here, $m_\odot = -26.77$ is the apparent visual magnitude of the Sun, and p_V is the visual geometric albedo of the nucleus (Tancredi *et al.* 2000).

A potential problem with the above procedure is that there is no guarantee that the photometry refers to the actual nucleus. Under many circumstances it is quite likely that the light emanates from both the nucleus and a surrounding cloud of dust. If so, the derived value of R_N will be an overestimate.

The problem is unavoidable, when we deal with real comets that experience activity in parts of their orbits. Usually, the visible activity (i.e., the coma) subsides as the comet moves away from the Sun, and it may even disappear completely. However, when this happens, the comet is mostly far away. The nucleus is a very faint object, and with the limited angular resolution of the imaging, even a sizeable dust cloud would appear starlike. This leaves open the possibility that such a cloud exists in the form of residual grains orbiting around the nucleus since the last period of activity. Such grains were positively identified in the Rosetta target comet (67P/Churyumov-Gerasimenko), as the probe explored the comet in the outer part of its orbit (Rotundi *et al.* 2015).

A second approach to the problem has been successfully pursued in many cases. This is to use the excellent angular resolution of space-borne telescopes and image the inner parts of the dust coma including the central condensation, where the nucleus dwells. For visual observations, the Hubble Space Telescope has been used (Lamy *et al.* 2004) taking advantage of the occasions, when comets pass relatively close to the Earth and the linear measure of each imaging pixel is minimal. By extrapolating the observed brightness of the

[1]This is a sphere, whose cross-section πR_N^2 equals that of the observed nucleus.

dust coma, one can estimate its contribution to the brightness of the central pixel, where the nucleus is situated. When subtracting this, one has to assume that the dust does not obscure the nucleus, but such obscuration would only occur in extremely active comets. Here, comets of low activity are preferentially selected. The apparent nuclear magnitude thus derived is potentially the most reliable of all remote data. However, the phase angle can be substantial, and the accuracy of the absolute magnitude may be limited by the uncertainty of the phase function as well as that of the extrapolated central dust brightness.

Especially when space telescopes are utilized, a detailed advance planning of the observations is of essence. Therefore, short-period comets are much easier to target than the long-period ones. While some observations of the latter have been made, they do not suffice for a study of the size distribution. In fact, the lack of data concerning the nuclear sizes of long-period comets is a serious problem, since it makes it much more difficult to constrain the masses of these comets and hence to estimate the total mass of their source population, the Oort Cloud. More or less, the same problem holds for the Halley Type comets, and the only group that has been well studied is the Jupiter Family.

2.1.2. *Thermal radiation*

This group was recently targeted for nuclear photometry in the mid-IR wavelength region using the Spitzer Space Telescope (Fernández *et al.* 2013). The wavelengths used were near 16, 22 and 24 μm, and thus the detector arrays received the thermal radiation of the comet nucleus and the surrounding dust. Coma corrections were applied, but the comets were at $r > 4$ AU to facilitate this procedure, and residual dust in the central pixel could not be ruled out. It was, however, concluded that the nuclear brightnesses were not overly affected by this.

As seen from Eq. (2.2), for visual photometry the absolute nuclear magnitude constrains the product of the visual albedo and the square of the radius. Since it is rarely possible to determine

the albedo independently (see Sec. 2.4), the estimate of the radius generally depends on an assumed albedo. The albedos that have been determined for Jupiter Family comets are very low, so an error in the assumed value can have a large effect. But fortunately, these albedos are rather similar, so using an average (0.04 is the most commonly used value) leads to radius estimates that should be fairly good in terms of internal consistency. However, for the mentioned mid-IR photometry the situation is better, since the quantity involved in that case is the emissivity (ϵ), which should have a higher relative accuracy than the albedo. This is because, as indicated by Kirchhoff's law, a low value of p_V implies a value close to unity for ϵ. Therefore, even if the error in p_V is quite significant in relative terms, the error in ϵ should be less important.

The constraint from mid-IR observations is obtained as follows. The thermal continuum flux density at frequency ν can be written

$$F_{\rm th}(\nu) = R_N^2 \epsilon(\nu) \frac{\Phi_{\rm th}}{\pi \Delta^2} \oint_{2\pi} B_\nu(T_s) d\omega, \tag{2.3}$$

where $\Phi_{\rm th}$ is the phase function at the relevant frequency. The integral over cometographic solid angle $d\omega$ is taken over the Sun-facing hemisphere of the nucleus. Here, $B_\nu(T_s)$ is the Planck function at frequency ν evaluated for the surface temperature T_s. The expression of $\Phi_{\rm th}$ in terms of thermal infrared magnitude $m_{\rm th}$ would be $\delta m_{\rm th} = -2.5 \log \Phi_{\rm th}$.

Accordingly, the solution for R_N requires a thermal model, from which a surface temperature map is obtained. Inherent in this are several additional parameters: at the very least, the Bond albedo, the bolometric emissivity, and the beaming factor that expresses the anisotropy of the thermal emission from a rough surface in terms of concentration to the solar direction. Many papers have been written on this topic, and recent works often use more sophisticated models than Eq. (2.3). A thorough discussion of the effects of roughness and thermal inertia is found in (Davidsson *et al.* 2015). While Fernández *et al.* (2013) were able to circumvent some of the modeling uncertainties by independent fitting of beaming parameters, these authors also noted the possible problems caused by the unknown

shapes — since the model uses a sphere, it may be inaccurate for non-spherical nuclei.

2.1.3. *Statistical analysis*

There is nowadays a good set of nuclear magnitudes and radii for Jupiter Family comets due to several surveys and data collection efforts. However, the data set of radii is not homogeneous. The close-up imaging provides the basic reference data. Visual absolute magnitudes in general are of varying quality depending on amount of data, orbital coverage, and observational techniques used. Only rarely have they been obtained together with simultaneous IR data for determination of radius and albedo together. The most prominent cases are comets 28P/Neujmin 1 and 49P/Arend-Rigaux, which are known since long ago for their very low activity. For a Table summarizing the basic physical properties of all such comets, see Sec. 2.4.2.

Judging from comparisons of the different data sets on nuclear radii, there are still large uncertainties over the actual radii of most individual comets. This became evident, when Fernández *et al.* (2013) made such comparisons between their own results and those from compilations of data from visual observations. While the average offsets between mid-IR and visual radii are not remarkable, the scatter around the 1:1 line is much larger than expected from non-sphericity or the spread in individual albedos.

Figure 2.1 shows the comparison of nuclear radii from Tancredi *et al.* (2006) and Fernández *et al.* (2013) for 38 comets as an illustration. The worst quality estimates according to the former paper are identified, and these do not stand out as the major culprits for the scatter. Moreover, looking separately at the seven comets of the two best quality classes, these do not stand out as especially good fits. The material does not include any comets visited by spacecraft, so there is no independent verification. It appears that the source of the scatter may be problems with the modeling involved or residual dust contributing to the nuclear brightness.

Both the visual and the mid-IR data sets have been used to derive the size distribution of Jupiter Family comet nuclei. The method used

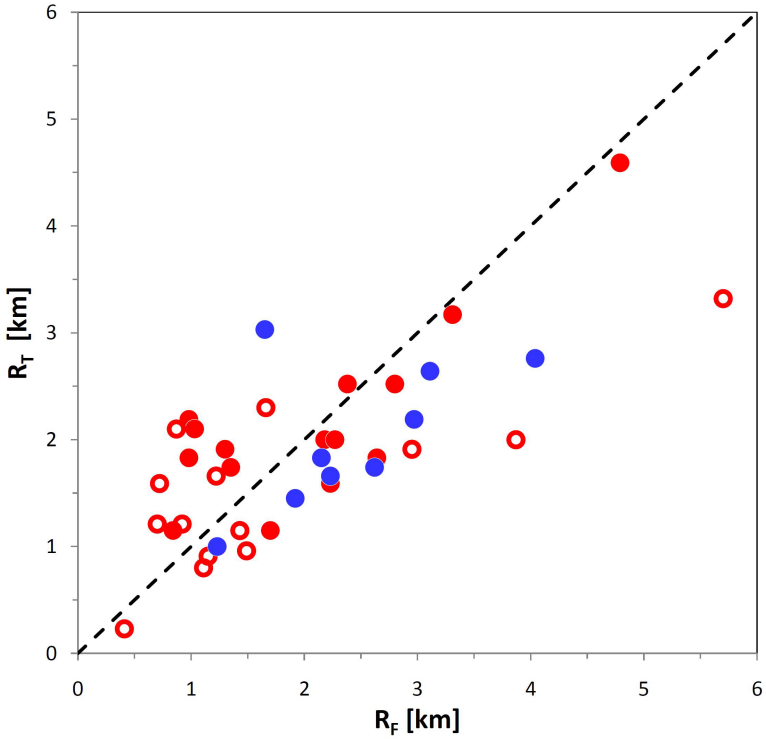


Fig. 2.1. Comparison of radii for Jupiter Family comets, determined independently by Tancredi *et al.* (2006) (R_T) and Fernández *et al.* (2013) (R_F). The dashed line indicates a 1:1 regression. Filled red circles show comets with R_T of quality class 3, and open red circles concern comets of quality class 4 (the lowest quality). The blue circles indicate quality class 1 or 2 (the best quality). Courtesy T. Wiśniowski.

is always basically the same. The cumulative distribution of nuclear radii is imagined to be well fitted by a power law, and the task is to fit the index (α) of this function according to

$$N(R > R_o) \propto R_o^{-\alpha}. \tag{2.4}$$

The straightforward way to do this is to plot $\log N$ versus $\log R_o$, fit a straight line to these data, and measure its slope.

However, there are problems with this method too. One problem is that the data set of nuclear radii may be inhomogeneous and of varying quality, causing errors in the horizontal location of the points to be fitted. Another problem is caused by incompleteness of the observed sample of comets, possibly causing a radius dependent discovery bias. A third problem is that, at the level where N is just a few (i.e., for the largest observed comets), the sampling of the parent distribution is obviously insufficient. Yet another, more fundamental

problem is that the parent distribution may not be an exact power law. If there was a very rich set of data points of high accuracy, this might not be a problem, since the actual shape including wiggles or waves could then be observed. But with a rather small amount of low-quality data, this may actually be a serious source of error.

Keeping in mind all these caveats, it is not very surprising that the record of published α values is rather discordant. In short, from visual data bases used during the last 20–25 years, the range of α extends from about 1.5 to 2.5, and there is no telling where the truth is. From their mid-IR survey, Fernández *et al.* (2013) made an ambitious search for the best power-law index (while acknowledging some evidence for deviations from a power law), which resulted in a preferred value of 1.9.

An important piece of indirect evidence was recently obtained from the NASA *New Horizons* mission. From studies of the impact craters on Pluto and Charon, Singer *et al.* (2015) concluded that the small projectiles (about km-sized) causing such craters have a size distribution with $\alpha \approx 2.3$. These projectiles belong to the trans-neptunian population, which includes the likely source of most Jupiter Family comets (see Sec. 5.4). Hence, there is good reason to assume that these comets start out with this value of the size distribution index. The index pertaining to the actual Jupiter Family comets, after some aging and erosion is probably not much different.

This is where the issue stands regarding the observational constraints on the size distribution of comets. Related issues concerning the number of comets in the solar system and their total mass, and the evolution of these quantities, will be discussed later (Chaps. 6 and 7).

2.2. Brightness and Gas/Dust Production

2.2.1. *Apparent total magnitude*

The most common measure of the activity of comets is found in the sets of visual magnitudes observed at different times during their apparitions. These represent the integrated brightness of the coma including the central condensation, to which the nucleus may

also contribute. The wavelength coverage is generally broad, and typically, the eye is the detector, unaided or using binoculars or the eyepiece of a telescope. Photographic and CCD data also exist in large quantities, but these typically fail to incorporate the whole coma. An advantage of these data sets is that the activity is often easily monitored, as the comet moves through the near-perihelion part of its orbit. For a periodic comet, it is also possible to compare its behavior on successive orbits and look for evolutionary changes.

However, there are also drawbacks. In contrast to other extended astronomical objects like gaseous nebulae or galaxies, which generally stay the same all the time, comets keep changing as a result of their orbital motions. They move across the sky, and they brighten or fade — sometimes quite rapidly. Therefore, in a sense, it is like observing different objects from month to month or night to night, and one cannot go back and reobserve to improve the accuracy of a measurement. This is particularly troublesome, since the comet magnitudes are susceptible to uncertainties caused by observational circumstances and observer-specific biases.

The reason for the problem is that the total magnitude should ideally encompass a rather large, diffuse object whose surface brightness varies strongly from the center to the outskirts. An important fraction of the light may be missed, if the comet is observed with too high magnification, or the outer coma is lost in the sky background. These effects would always lead to underestimates of the integrated brightness, but in reality there are other problems too, affecting the measurements by most individual observers. Hence, even if lots of data are available (which is not often the case), deriving an accurate light curve in terms of total magnitude versus time during a cometary apparition is far from being a trivial task.

Nonetheless, these data provide the only way to determine quantitatively, how complete the observational record is, and to set a standard for counting the numbers of comets in different populations. This is because it is essentially the total magnitude of a comet that determines its chances to be detected or discovered. Consequently, the data have to be carefully selected and seriously discussed. The light received remotely from a comet coma consists

of two parts: sunlight scattered from the dust grains, and what is traditionally interpreted as fluorescent emission from a few radicals (molecular fragments) populating a region of radius $\sim 10^5$ km around the nucleus. The most prominent of these are C_2, C_3 and CN, each shining in its particular wavelength bands. The so-called Swan bands of C_2 are often dominant. The ratio between the dust and gas contributions is variable, but generally they are of similar magnitude.

2.2.2. *Absolute and heliocentric magnitudes*

It is obvious that the total apparent magnitude of a comet should depend on its distances from the Sun and the Earth. Correcting for this dependence will lead to an *absolute magnitude*, revealing the intrinsic brightness of the comet. As expected, the flux received on Earth turns out in most cases to fall off as the inverse square of the geocentric distance, corresponding to a term $5 \log \Delta$ in the expression for the total magnitude like in Eq. (2.1) for the nuclear magnitude. However, for comets approaching the Earth within 1 AU it may happen that the angular extent of the coma increases so much that the photometry suffers from the same problem as when using a too high magnification. The apparent brightness then varies with Δ more slowly than the second power. This is sometimes called the *Delta effect*.

The dependence on the heliocentric distance is less obvious. A standard correction that is often applied to the total magnitude uses a term $10 \log r$, which corresponds to a fall-off of the brightness as r^{-4}. This has some support in the following argument. We consider the comet to be so close to the Sun that the entire solar energy input to the nucleus (decreasing as r^{-2}) goes into sublimation of ice with associated outflow of gas and dust. Within a short time, the parent molecules of the luminous radicals get dissociated, so the amount of dust grains and radicals in the visible coma also drops as r^{-2}. But the flux of radiation emanating from a given amount of such material is proportional to the flux of sunlight that it receives — at least in the case of fluorescent emission by the gas. Thus we have two factors r^{-2} that should be multiplied.

The resulting formula for the total magnitude (m_1) of a comet is:

$$m_1 = H_{10} + 10 \log r + 5 \log \Delta, \tag{2.5}$$

where r and Δ have to be expressed in AU. Formally, this would mean that H_{10} is a *total, absolute magnitude* of the comet, expressing its intrinsic brightness reduced to a distance of 1 AU from both the Sun and the Earth. Contrary to Eq. (2.1), there is no phase term in Eq. (2.5). This is natural, because a phase effect would only be present in the dust contribution, and it would be difficult to characterize without detailed knowledge of the grain properties.

However, the H_{10} magnitudes are of dubious value. On the one hand, the argument for using the r^{-4} dependence is weak at best. Several points can be made. The use of r^{-2} for the production rate of gas and dust is a serious oversimplification for any individual comet, as evidenced over and over again by the shapes of the actual, accurate light curves and the direct observations of gas production curves. After a lot of progress during the last decades and, in particular, the exploration of comet 67P by Rosetta, we have some insight into the mechanisms behind the real production curves, and the use of a r^{-2} law should now be regarded as obsolete.

On the other hand, the second factor r^{-2} is suspect too for different reasons. The dust contribution to the coma brightness must depend on the scattering phase function of the optically dominant grains. When comets are observed near opposition, this is not very important, but many times bright comets are observed not very far from solar conjunction, and their brightness may be enhanced by a tendency for forward scattering by the smaller grains. The gas contribution, according to recent Rosetta results (Bodewits *et al.* 2016) may not be realistically modeled as pure fluorescence, since the excitation of the radicals is often dominated by electron impact. Thus, the question remains, whether the number density of electrons in a typical coma varies as r^{-2} or not.

Moreover, the bulk of experience on comet light curves has shown that the more general formula

$$m_1 = H + 2.5n \log r + 5 \log \Delta \tag{2.6}$$

may rather be used with $n \neq 4$. In this case, H is still an absolute magnitude referring to the total brightness of the comet extrapolated or interpolated to $r = \Delta = 1\,\mathrm{AU}$. Of course, the predictive power of Eq. (2.6) is limited by the need to determine the *photometric index* n individually for each comet or estimate it by some other means.

There is another quantity that is of more direct use, namely, the *heliocentric magnitude* (m_h):

$$m_1 = m_h + 5 \log \Delta, \tag{2.7}$$

which measures the brightness of the comet at the respective position in its orbit, only corrected for the geocentric distance. This should somehow be related to the gas production rate of the comet and may therefore be used as an indicator of the gas production curve. In particular, Jorda *et al.* (1992) showed that there exists a correlation between the heliocentric magnitudes of comets and their measured H_2O production rates ($Q[H_2O]$) with a regression formula

$$\log Q[H_2O] = 30.74 - 0.24 m_h. \tag{2.8}$$

The error bars on the numerical parameters are small. However, the sample of comets was small (13 objects), and water production rates were only generally known from the OH lines at 18 cm observed with the Nançay radio telescope. Equation (2.8) implies that the water production rate varies as the 0.6 power of the intrinsic coma brightness, in rather good agreement with the value of 0.5 suggested by the above theoretical argument.

Naturally, to provide the outgassing rates of gas molecules and the release rates of dust, direct observations are much superior to the visual magnitudes. The problem with the gaseous species, i.e., the parent molecules of nuclear origin that give rise to the observed radicals, is the opacity of the Earth's atmosphere for most of the critical transitions at infrared and sub-mm wavelengths. To this comes the limited access to space telescopes with the necessary cryogenic equipment, especially for monitoring purposes. Monitoring can often be made for the radicals by ground-based narrow-band photometry or spectroscopy, but it is often difficult to relate the observed production rates to those of specific nuclear parents. In practice, the production

rates of the radicals are often translated into water production rates by a simple, empirical correlation coefficient.

2.2.3. *Dust production rates*

As regards the dust, the situation is even more complex. The coma can be imaged using filters that transmit the continuum with minimal contamination by gas emission, so that the flux is attributed to the totality of dust grains within the field of view through scattering of sunlight. In this case, the $Af\rho$ parameter (A'Hearn *et al.* 1984) is generally used to constrain the integrated cross-section of all these grains via the filling factor f.[2] This parameter is easily derived from the observed quantities, where ρ is the radius of the field of view at the comet. A good estimate of the grain albedo A thus allows us to find the integral of the grain area over the entire population. Clearly, this is somehow related to the dust mass production rate, but conclusions about the latter are hampered by strong model dependence.

There are two features that would need to be taken into account but cannot be straightforwardly assumed. One is the *dust size distribution* (DSD), and the other is the expansion velocity of the dust as a function of grain size. Observational estimates of the DSD are generally made by analyzing the brightness distribution over cometary dust tails, in either visible or infrared light. As a model for the differential DSD, a power law with index β is used for the radius interval under consideration. Here, the equation used is analogous to Eq. (2.4) in Sec. 2.1.3 with $\beta = \alpha + 1$.

This issue has been reviewed by Fulle (2004) with the following conclusions. Oversimplification of the dynamical model used is extremely risky, as was seen during the ESA/Giotto flyby of comet 1P/Halley in 1986, when the spacecraft was hit by a grain far too large (its mass was $\sim$1 g) to be compatible with the advance predictions. Nowadays, Monte Carlo methods are preferred to invert the brightness distribution with the use of a full description of the

[2]This is defined to be the fraction of the solid angle in the field of view, where the line of sight crosses a grain.

dust dynamics. Resulting values of β tend to be in the range between 3 and 4, implying that the total dust mass is dominated by the largest grains, which may actually be meter-sized boulders. Even the scattered, visual light from parts of the dust tails may involve large chunks to an extent that previously seemed hard to imagine.

The actual dust mass loss rate will of course depend on the maximum size of the ejected particles. In comet 1P/Halley, during the Giotto flyby, the dust impact detection system (DIDSY) detected the mentioned 1 g particle, thus showing that chunks of at least this size were ejected. Thus, the ratio of dust to gas mass loss rate from the nucleus — the so-called *dust/gas ratio* (DGR) — was at least of order unity, which came as a surprise in view of earlier, much lower estimates. Since that time, the signatures of large coma particles in some comets have been measured by both radar at cm wavelengths and radio observations at mm wavelengths. High dust mass loss rates have thus been derived, implying large DGRs in the observed comets. In addition, similar conclusions have come from studies of *dust trails* associated with several comets, observed at infrared wavelengths. These trails consist of large grains that have left their parent comet at low speeds and slowly drifted away from it counter to the direction of orbital motion. Sykes and Walker (1992) concluded that a mass ratio of about 3 between ejected refractories and gas was indicated by the dust trail observations.

Recently, dust measurements by Rosetta instruments in the coma of comet 67P were analyzed (Rotundi *et al.* 2015), revealing that at 3.4–3.6 AU inbound, the ejected grains had diameters up to $\sim$2 cm, and the larger chunks in orbit around the nucleus since a time near the preceding perihelion were sometimes meter-sized. These estimates translate into a DGR of $\simeq$3 near perihelion and $\simeq$4 for the material ejected in the mentioned part of the comet orbit. The DWR (dust/water ratio) would be $\simeq$6, but Rosetta/ROSINA data for CO and CO_2 had shown these molecules to contribute $\sim$50% in mass relative to H_2O (Hässig *et al.* 2015).

These results may imply a paradigm shift in comet science. Before the Halley exploration, comet nuclei were described in popular wording as "dirty snowballs", and the "dirt" — i.e., the

dust — was generally considered a minor contamination. Afterwards, the term "snowy dirtballs" has occasionally been used in a somewhat provocative way, and this may be more to the point. But two issues should be kept in mind. First, the number of individual comets for which large DGRs seem confirmed remains small, and there is no way yet to say how the DGR varies between different kinds — say, between periodic comets and fresh, long-period comets. Second, the surface layers of comet 67P are heavily processed as a result of insolation, ice sublimation and grain redeposition (so-called *airfall* — see Sec. 4.2.1) during many orbits, and the same is likely true of most periodic comets. This makes it risky to conclude that a large DGR of the material currently leaving the surface can be interpreted as a large ratio of refractories to ice in the bulk of the nucleus.

2.3. Light Curves and Gas Production Curves

One of the most obvious ways to study the nature of comet nuclei is to observe in detail, how the activity of the comet responds to the changing insolation, as the comet approaches the Sun before perihelion and recedes afterwards. This was the basic, direct goal of the Rosetta mission: to learn "how comets work" by escorting comet 67P from far out through perihelion and thereafter, observing from close range how the activity developed and the nucleus surface changed. As often emphasized, the Rosetta results are truly remarkable and deeply impressive, but strictly speaking, they refer to one comet only. The major scientific harvest of Rosetta calls for a way to generalize the findings, to put 67P into a wider context, and thus to understand much better, not only how this comet works but how all comets work. An Earth-orbiting telescope with UV and IR detectors primarily aimed to monitor the activity variations of many other comets is an example of a facility that could provide a good match to Rosetta but has not yet been realized.

Therefore, we have to make the best use of what we have, and this means in particular to study comet light curves along with evidence from imaging and measurements of production rates of gas species and dust grains. Figure 2.2 illustrates, as an example, the heliocentric

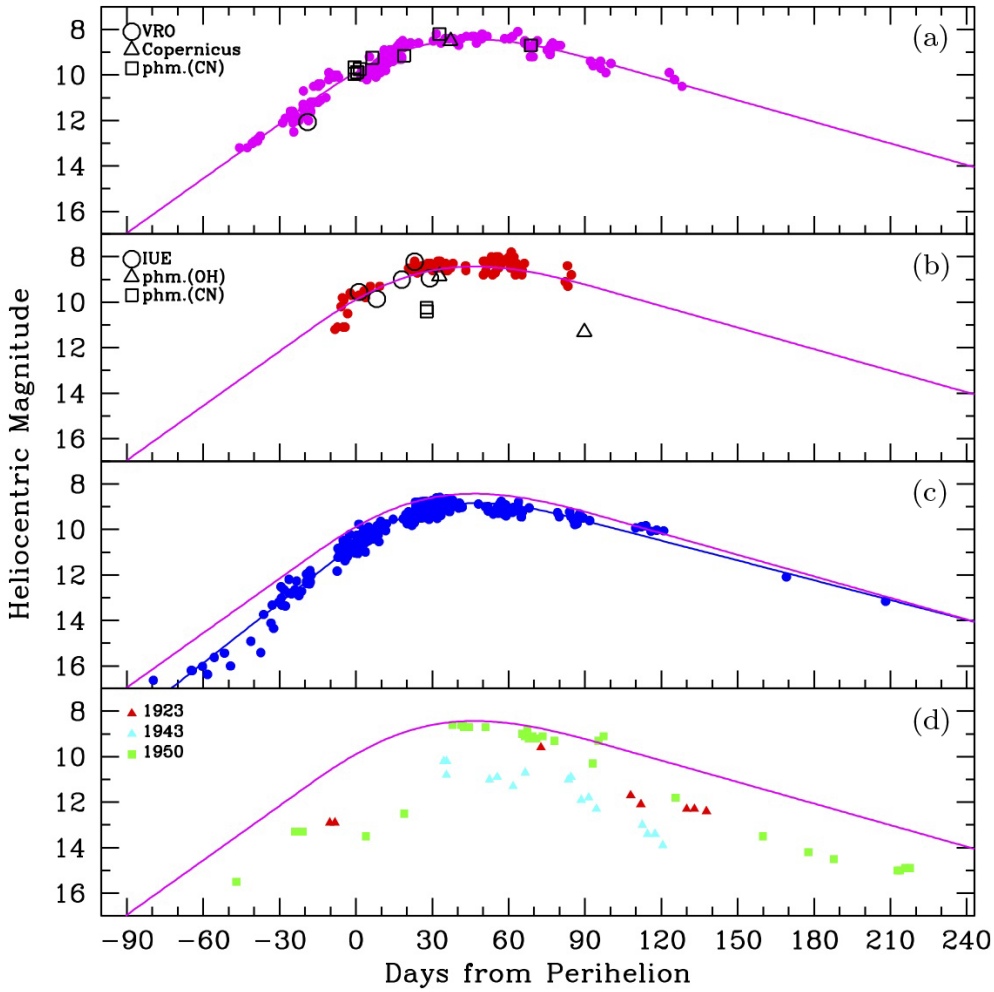


Fig. 2.2. Observed light curve for comet 6P/d'Arrest as observed in 1976 (panel a), 1982 (panel b), 1995 (panel c) and during three earlier apparitions (panel d). The magenta and blue curves show polynomial fits to $m_h(t)$ in 1976 and 1995, respectively. Reprinted from Szutowicz, S. and Rickman, H., *Icarus* **185**, 223–243 (2006), with permission from Elsevier.

magnitude of comet 6P/d'Arrest during three favorable apparitions together with some direct observations of its gas production rates. Similar data are available for many comets although rarely of comparable quality. As seen from Eq. (2.8), the plotted curve directly translates into the variation of the log of the H_2O production rate with time.

2.3.1. *Isothermal model*

To see how this relates to the properties of the nucleus, let us first consider a very simple physical model. We assume that the nucleus is spherical and that its surface is smooth and covered with H_2O

ice. Placing it at a certain distance (r) from the Sun, it is clear that different local elements of this surface receive different fluxes of solar radiation, corresponding to the local zenith distance (ζ) of the Sun. However, we disregard this circumstance and use a common solar zenith distance all around the nucleus. The projection factor that governs the energy absorption rate by the surface is $\cos\zeta$, and its spherical average is 1/4. This is sometimes inaccurately called the rapid rotator approximation, because one may imagine a nucleus spinning so fast that its thermal inertia causes the whole surface to have the same temperature (T). Calculating this temperature, one should then use the average of $\cos\zeta$. Another, more accurate name for the procedure is hence the *isothermal approximation*.

The energy absorption rate should be balanced by the loss rates. One of these is the thermal radiation of the surface, which is found from the Stephan-Boltzmann law. Another is the latent heat (H) consumed by the sublimating H_2O molecules, and this must be combined with an expression for the sublimation rate (Z) — i.e., the flux of molecules leaving the surface — as a function of temperature. The third loss rate (either positive or negative) would correspond to the heat exchange between the surface and the sub-surface layer, but in the isothermal approximation one also neglects the vertical temperature gradient, so this term vanishes.

According to the simplest gas-kinetic sublimation theory, the flux of escaping molecules can be written in terms of the product of the number density in a saturated gas (n_s) and the average velocity (v_{th}) of the molecules. From this, one derives

$$Z(T) = \frac{p_s(T)}{\sqrt{2\pi mkT}}, \tag{2.9}$$

where p_s is the saturation pressure of the gas, m is the molecular mass and k is Boltzmann's constant. To a good approximation, p_s is an exponentially increasing function of T, and hence, $Z(T)$ shares the same qualitative behavior. Using a constant value for H, the energy balance equation for the surface is

$$\frac{1}{4}F_\odot r^{-2}(1 - A_v) = \epsilon\sigma T^4 + H \cdot Z(T), \tag{2.10}$$

where $F_\odot$ is the solar constant,[3] A_v is the hemispherical visual albedo and ϵ is the thermal emissivity of the surface. Using appropriate values for A_v and ϵ, and taking the thermophysical data (H and p_s) for H_2O, one can easily solve for $Z(r)$, and the same can be done for other molecules too, if the comet should be composed of other ices.

Examples of such solutions are seen in Fig. 2.3. The exponential behavior of $p_s(T)$ is common to all molecules and leads to a common feature of the curves. Close enough to the Sun, where T is high enough, the sublimation loss term dominates over the thermal radiation. This means that the solar energy flux goes essentially into sublimation, and the sublimation rates tend to fall off as r^{-2}. On the other hand, when the comet is far from the Sun, the thermal radiation dominates over sublimation. Consequently, T falls off nearly as $r^{-1/2}$, causing an exponential drop of the sublimation rate. The transition

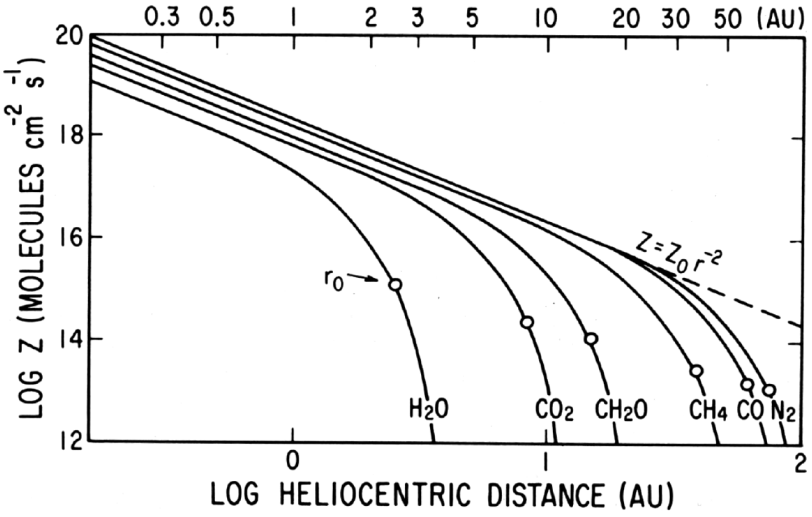


Fig. 2.3. Sublimation rates for a selection of cosmically important molecules versus heliocentric distance, plotted in a log–log diagram. A dashed line indicates a fall-off as the inverse square of the distance. The open circles marked r_0 indicate the knees, where the sublimation and thermal radiation heat losses are equal. See Delsemme (1982). From *Comets*, edited by Laurel L. Wilkening.

[3]The solar constant is the amount of energy passing per unit time through a unit surface perpendicular to the solar direction due to solar radiation at all wavelengths, at a distance of 1 AU from the Sun.

between these two regimes is marked by a knee on the curve, and the location of this knee depends on how volatile the substance is, i.e., how large its saturation pressure is at a reference temperature.

As one can see, H_2O is less volatile than all the other selected, cosmically abundant substances. A comet made of H_2O will start sublimating appreciably at a distance around 2–3 AU from the Sun, while comets made of the other species would become active much further away. Here is the reason why comets were considered to be made mostly of water ice long before this could be verified by spectra or in situ exploration. Their activity tends, in a statistical sense, to commence and subside at heliocentric distances close to the knee of the H_2O curve.

Thus, one may imagine that this H_2O curve may yield at least a crude fit to the observed H_2O production curves of different comets. For each comet, it would be possible to multiply $Z(r)$ by the surface area of a sphere with the same radius as the comet nucleus to get an idealized production curve for a fully active nucleus. This would be shifted by some amount from the observed curve, and the shift could be interpreted to mean that the activity is limited to some constant fraction of the surface. Hence, the fitting would mean to find the relevant value of the active fraction. However, reality is not so simple.

2.3.2. *Seasonal variation*

The thermal model of Eq. (2.10) is of course extremely crude and may only be useful to arrive at general conclusions like the importance of H_2O. However, for interpreting real light curves, it is inadequate. This is obvious from Fig. 2.2 already at first sight. As seen, the light curve of comet 6P/d'Arrest — and, therefore, the H_2O production rate according to Eq. (2.8) — peaks about 40 days after perihelion, while Eq. (2.10) predicts Z to be a monotonously decreasing function of r, and thus the light curve has to peak at perihelion and to rise and fall on either side in a symmetric way. Even if comet d'Arrest is a rather extreme example of perihelion asymmetry, the phenomenon is quite general, and it would be hard to find any comet exhibiting a light curve in full agreement with Eq. (2.10).

How should this fact be explained? Which underlying assumption shall we abandon? Obviously, comet nuclei cannot be absolutely spherical, and their non-sphericity is verified by all the images acquired to date. But it is not obvious that such large disagreements as seen can arise from shape effects only. The idea of a thermal lag, whereby the surface is preferentially cooled by heat conduction toward the interior before perihelion, can also be discarded, since the light curve peaks are often found on the pre-perihelion orbital branch. Another obvious conclusion is that real nuclei cannot be isothermal. However, by simply introducing the temperature contrasts of a day-night cycle, one will not solve the problem of perihelion asymmetry, even though the temperature maps may become more realistic.

A different improvement is much more promising and also verified by imaging: replace the isotropic, more or less icy surface by one where ice exists only locally if at all, and gas production from the remaining area is quenched by an overlying, refractory or ice-depleted layer. In this case, depending on the orientation of the spin axis, seasonal variation in the insolation of the most active (least quenched) spots may cause a strong perihelion asymmetry in the light curve, leading to a maximum either before or after perihelion. Seasonal variations in comet activity were first considered by Paul Weissman (1986).

It is clear that the sources of comet activity are to some extent local and thus in general subject to seasonal effects. Some models of comet nuclei have involved *active spots* of constant size with a freely sublimating icy surface, while the rest is completely inactive. The observed perihelion asymmetry may then be explained by a proper choice of spin axis and spot latitudes. The fit to the observed gas production curve then consists of choosing the best areas of the active spots, and the sum of these areas divided by the total surface area of the nucleus then yields the active fraction.

However, reality is certainly more complicated. On the one hand, the black and white picture of such models with surface elements that are either 0% or 100% active is oversimplified, as shown by the Deep Impact and Rosetta missions (Sec. 1.5). Most of the gas production of the target comets comes from regions with activity levels larger

than zero but less than 100%, as one may expect for ice sublimation at shallow depths below the surface, involving partial quenching of the outflow.

On the other hand, given that we can speak of such an activity level as a continuous variable between 0 and 1, there is no reason to think that it should be everywhere the same or that it should stay the same as time proceeds at any given place. Many surface changes on comet 67P during its 2015 apparition have been recorded by the Rosetta/OSIRIS imaging, indicating likely changes in activity level as well. Thus, it appears that comet light curves and gas production curves may reflect a complex interplay of surface phenomena in addition to the fundamental variation of the ice sublimation rate due to the varying heliocentric distance. To some extent, these curves may hence be beyond the predictive capability of simple models, but the bulk features should clearly be understood. For instance, the observed perihelion asymmetries are often quite repeatable from one orbit to the next and should contain important information about the activity distribution and nucleus spin. Other features of similar importance include the width of the light curve peak, and the fall-off slopes on either side as indicated by the photometric index — see Eq. (2.6) in Sec. 2.2.2.

2.3.3. *Remote activity*

One more feature remains to be described. This is the gas-producing activity at large heliocentric distances, which may call for other explanations than H_2O ice sublimation.[4] There are many examples among both short-period and long-period comets. The latter include at least one comet that made the newspaper headlines, namely, C/1973 E1 (Kohoutek). At discovery on 7 March, 1973, this comet was more than 4 AU from the Sun, and its perihelion distance was only 0.14 AU. Considering the heliocentric distance, the discovery magnitude of 15 was quite bright, and extrapolation to perihelion by a standard photometric index led to the prediction of a magnificent

[4]Other species than H_2O can be important close to the Sun as well, as shown by the EPOXI observations of comet 103P/Hartley 2 (see Sec. 1.5.4).

display, which triggered the use of the term "comet of the century" in news media. When comet Kohoutek came close to perihelion in December and January, it was a respectable object reported to be of second magnitude with a long tail, but this did not live up to the early expectations. As a result, most people who remember comet Kohoutek think of it as a failure.

Although we cannot tell for sure what happened to comet Kohoutek, in retrospect its brightness evolution does not appear enigmatic. The activity observed at discovery may have been caused by the release of species more volatile than H_2O. If so, there was no reason for a very rapid increase as the comet approached the Sun. At some point, closer to the Sun, the H_2O outgassing likely became dominant, but the brightness then took off from a lower level than predicted. This behavior might have been similar to the one observed in the famous comet C/1995 O1 (Hale-Bopp) as illustrated in Fig. 2.5 below, even though there is an important difference between the orbits of the two comets: comet Kohoutek was a new Oort Cloud comet, and Hale-Bopp was a returning long-period comet.[5]

When Jupiter Family comets pass through the aphelion part of their orbits at heliocentric distances $\simeq$5–6 AU, they often appear totally inactive but not always. For instance, Licandro *et al.* (2000) reported on a program to observe such comets far from the Sun to determine the photometric cross sections of their nuclei, but seven out of the 18 targets were deemed active, whereof six at $r >$ 4 AU. Figure 2.4 shows two examples of the images obtained: comet 37P/Forbes showing no activity at $r = 3.6$ AU and comet 65P/Gunn exhibiting a large coma at $r = 4.3$ AU. Other observing programs too have reached similar conclusions. In their above-mentioned mid-IR photometric study of Jupiter Family comets (see Sec. 2.1.2), Fernández *et al.* (2013) found extended dust emission in 35 out of 89 detected comets — all at $r > 4$ AU.

The question remains, what does it mean that many comets produce dust comae at large heliocentric distance? Is the dust

[5]The concepts of new and returning comets are explained in Sec. 5.1.

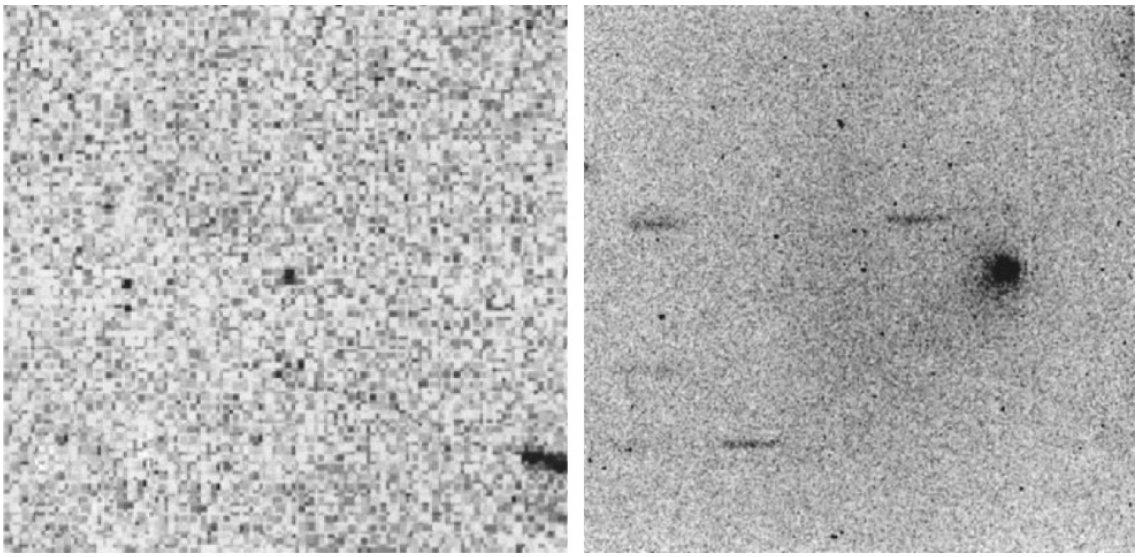


Fig. 2.4. CCD images of comets 37P/Forbes (left) and 65P/Gunn (right), obtained at the ESO/La Silla and Pic du Midi observatories, respectively. Reprinted from Licandro, J. *et al.*, *Icarus* **147**, 161–179 (2000), with permission from Elsevier.

production necessarily a direct consequence of gaseous outflow? If so, how much gas is required? Quite possibly, there may not be a unique relation between the strength of the gas outflow and the amount of dust particles forming the coma. Too little is yet known about how the refractories are lifted off the surface of the nucleus and how the grains are derived from what is likely a coherent, refractory matrix. In any case it is clear that some gas is needed to accelerate the grains away from the nucleus into the observed large, roundish cloud.

The amount of data on gas production rates of distant comets is quite small. Figure 2.5 shows the only impressive data set, obtained from radio astronomical observations of C/1995 O1 (Hale-Bopp) by Biver *et al.* (2002). Of the molecules included, CO was clearly the dominant one, when the comet was at $r > 3.5$ AU before perihelion and at $r > 3$ AU afterward. The $Q(CO)$ production curve at such distances is much shallower than the corresponding curve for $Q(H_2O)$, but closer to perihelion, the CO/H_2O ratio was more constant. This so-called Christmas tree diagram contains lots of information on how various gases may have been produced from this comet in different parts of its orbit. At this point, let us note that one likely, major molecule is missing for lack of observations, namely, CO_2.

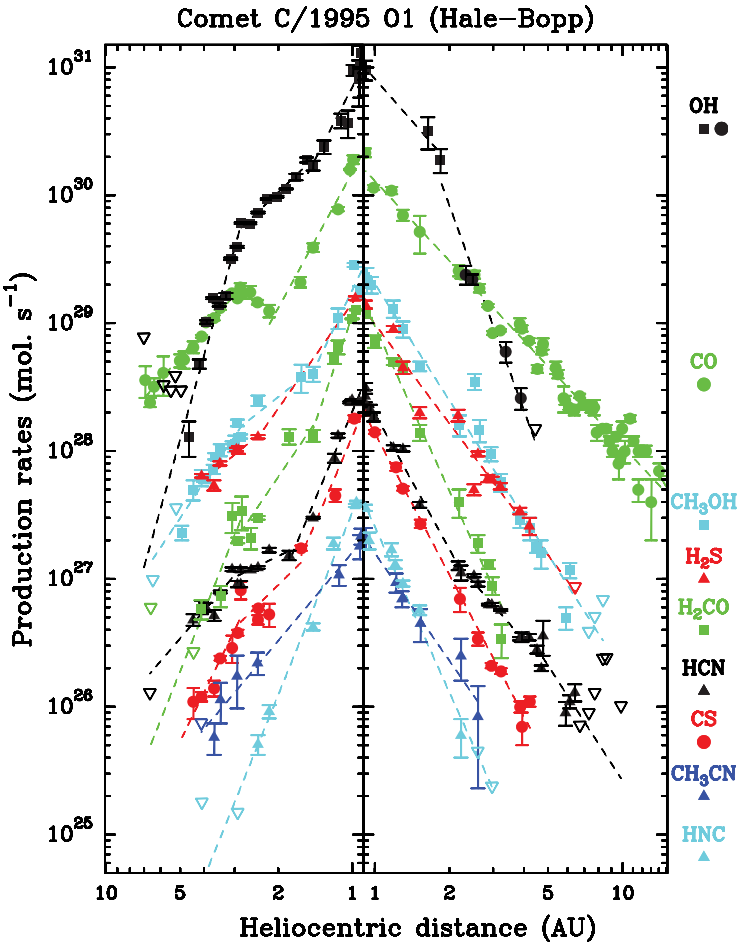


Fig. 2.5. Observed production rates of different molecules in comet C/1995 O1 (Hale-Bopp), plotted versus heliocentric distance in a log-log diagram. The perihelion passage at $q = 0.914$ AU is marked by the vertical line. Pre-perihelion observations are to the left and post-perihelion observations to the right. From Biver, N. *et al.*, *Earth, Moon, and Planets* **90**, 5–14 (2002), © Kluwer Academic Publishers. With permission of Springer.

The CO molecule has been observed in other distant comets too — notably, comet 29P/Schwassmann-Wachmann 1. This comet orbits around the Sun at distances of about 6 AU and is known to exhibit a persistent grain coma. In this case the word "grain" should be preferred to "dust", because the grains are likely composed partly of H_2O ice. This may in fact be a general phenomenon in comets too far from the Sun for such grains to sublimate at an appreciable rate, while gases like CO may drag them into the coma. Comet 29P is also famous for its intermittent outbursts when, apparently, huge quantities of grains are suddenly ejected from the nucleus and the comet brightens up by five magnitudes or more. The reason for these

outbursts remains unknown, while there are some clues as to the mechanism of CO outgassing, to be discussed in Sec. 2.6.6.

2.4. Albedo and Activity

A good size determination for a comet nucleus in principle opens up for estimates of two important physical parameters — the albedo and the level of activity. The albedo is then found from Eq. (2.2), using the radius and the visual absolute magnitude. Of course, a good size determination then means a direct one that does not depend on an assumed albedo! This limits the sample to a handful of comets that have been visited by spacecraft or were subject to observing campaigns involving both visual photometry and thermal radiometry, allowing determination of both radius and visual albedo simultaneously. While spacecraft targets can have any amount of activity, the second method is limited to comets with minimal activity, so that the observations approximately refer to the bare nucleus.

In Sec. 2.3 we saw how the word activity can be used to describe the relation between the actual water production rate of a comet and the production rate corresponding to a model surface with the same size as the nucleus, made up by H_2O ice under some assumptions. Some such assumptions were discussed and deemed not to be realistic. This means that the derived parameter, which may be called *active fraction* or *activity level*, is model dependent. It is linked to a physical model for the sublimation rate from the surface — a model that should be chosen so as to imitate the observed H_2O production curve as closely as possible. However, even if a very good fit is found, there is no guarantee that the chosen model is a true description of the nucleus.

In other words, the activity level has no prescribed physical interpretation. Quenching of the gas flow is a reasonable expectation in the case of H_2O production beneath the surface, but this holds only for local outgassing. The global activity level of an entire nucleus depends on other things as well, as we shall see for the case of comet 67P below. The only message conveyed by this parameter is that the

comet produces a certain fraction of the predicted H_2O flux from an idealized standard nucleus with ice on its surface.

2.4.1. *Activity and thermal modeling*

Concerning the model that underlies the definition of activity, there are two ways to go. If the comet is very well observed and its H_2O production curve is known in some detail, and if its nucleus is very well imaged so that a digital terrain model of reasonable quality is available, one may pursue the thermal modeling to a very high degree of sophistication. Within the framework of such a model, the resulting activity parameter has a well-defined physical meaning. It thus tells something useful about the comet in question, but it cannot be generalized. However, for most comets, the observational constraints are far from adequate to develop such a model. We must then use a suitably crude model that is reasonably realistic and applicable to the sparse data typically available for the average comet.

The isothermal model is too extreme by underestimating the sublimation rate, unless the comet is so close to the Sun that the insolation is anyway balanced by sublimation and thermal radiation is unimportant. Most Jupiter Family comets do not come close enough for this to hold true. For increased realism, the simplest model (Weissman and Kieffer 1981; Rickman and Froeschlé 1983a) introduces the instantaneous value of $\cos\zeta$ and features another term in the energy balance, corresponding to the heat exchange between the surface and the interior:

$$\cos\zeta F_\odot r^{-2}(1-A_v) = \epsilon\sigma T^4 + H\cdot Z(T) - K\left(\frac{\partial T}{\partial z}\right)_0, \qquad (2.11)$$

where K is the thermal conductivity at the surface, and the partial derivative expresses the temperature gradient with respect to depth z counted downward from the surface. This is to be treated as a boundary condition to the heat diffusion equation, which is solved numerically for the temperature $T(z,t)$ using also an interior boundary condition. The latter is usually taken to be isolating, i.e., the temperature gradient is zero at the lower boundary of the computational grid. However, when solving explicitly for the

diurnal heat flow that occurs in a very thin surface layer, it may be advantageous to use a non-zero thermal gradient at the bottom, resulting from the orbital heat flow that extends to larger depths.

These equations may be solved for local spots at different latitudes on the nucleus surface. For the diurnal heat flow, the factor $\cos\zeta$ then depends on the orientation of the nucleus spin axis, the latitude and the rotational phase. Seasonal effects generally appear, if attention is focused on certain latitudes. This may thus be useful for investigating comets whose gas production curves have perihelion asymmetry. However, for the definition of activity in the general case, it is better to neglect seasonal effects by assuming the spin axis to be at right angles to the orbital plane.

For an arbitrary spin axis orientation, the poles are often close to the orbital plane, and it is worth considering a model predicting the water flux from a pole facing the Sun as an extreme case (opposite to the isothermal one). In this case, Eq. (2.10) can be used with only one modification — the factor $\frac{1}{4}$ is replaced by 1. In Fig. 2.6 the two extreme models are compared. We see that the difference

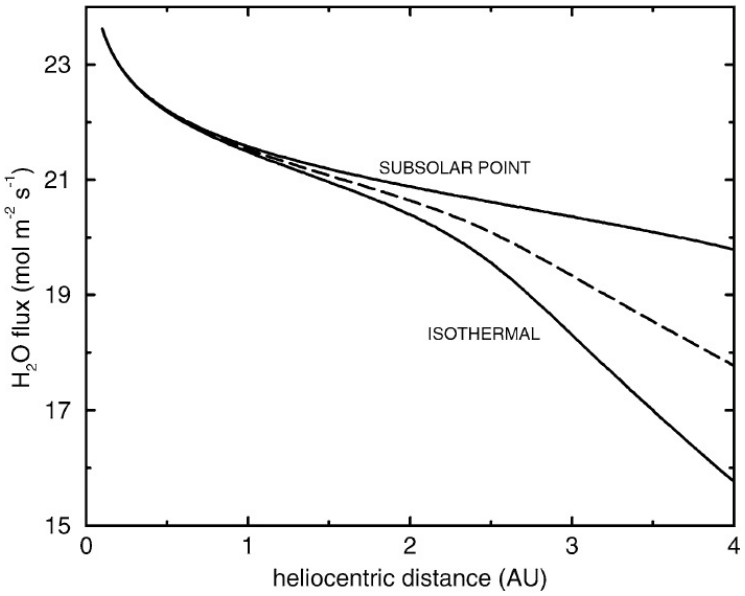


Fig. 2.6. H_2O sublimation flux versus heliocentric distance for an icy surface, computed for a pole facing the Sun (subsolar point) and the average surface of an isothermal sphere, represented by solid curves. The dashed curve is an average of these two models. Reprinted from Tancredi, G. *et al.*, *Icarus* **182**, 527–549 (2006), with permission from Elsevier.

in H_2O flux is moderate or small for $r < 2\,\mathrm{AU}$, while it grows to orders of magnitude for $r > 3\,\mathrm{AU}$. Since the heat diffusion model just described will always be intermediate between the two extremes, we may replace it by a simple average between the latter, if we restrict our attention to gas production rates observed at $r < 2\,\mathrm{AU}$ (also shown in the figure).

Tancredi *et al.* (2006) used this average to derive what they called active fractions for a sample of Jupiter Family comets with estimated water production rates from individual observations (single or in small groups) made at $r < 2\,\mathrm{AU}$ with the aid of their own values for the nuclear radii (see Sec. 2.1.3). These fractions f came from the identity

$$Q(H_2O) = 4\pi R_N^2 \cdot fZ, \tag{2.12}$$

where Z is the average sublimation flux at the heliocentric distance of the observation. Their results are shown in Fig. 2.7. Here we prefer the term activity rather than active fraction to avoid giving the impression of a 100% active area on an otherwise inactive surface.

2.4.2. *Statistical trends*

The most remarkable feature of the above diagram is that the values of f span the whole range from very close to zero up to more than 1. The largest values (and also the longest "error bars") are found for sub-km nuclei, and for $R_N > 2\,\mathrm{km}$ all the activities are less than 0.4. Only three comets have $R_N > 3\,\mathrm{km}$, and all of these have extremely low activities. It hence appears that there may be a correlation such that smaller comets have higher activities, and models of the physical evolution of comets should probably aim to explain this phenomenon.

While most comets that do not belong to the lowest quality class for nuclear radius have $f < 0.2$ (Tancredi *et al.* 2006), there are also a few such comets with much higher activities. Those for which f is consistently larger than 0.8 are: 21P/Giacobini-Zinner, 22P/Kopff and 46P/Wirtanen. The observational material seems solid, except that the radius of 21P (1.0 km) has a rather large error bar, being of quality class 3. There is hence no reason to doubt the existence of

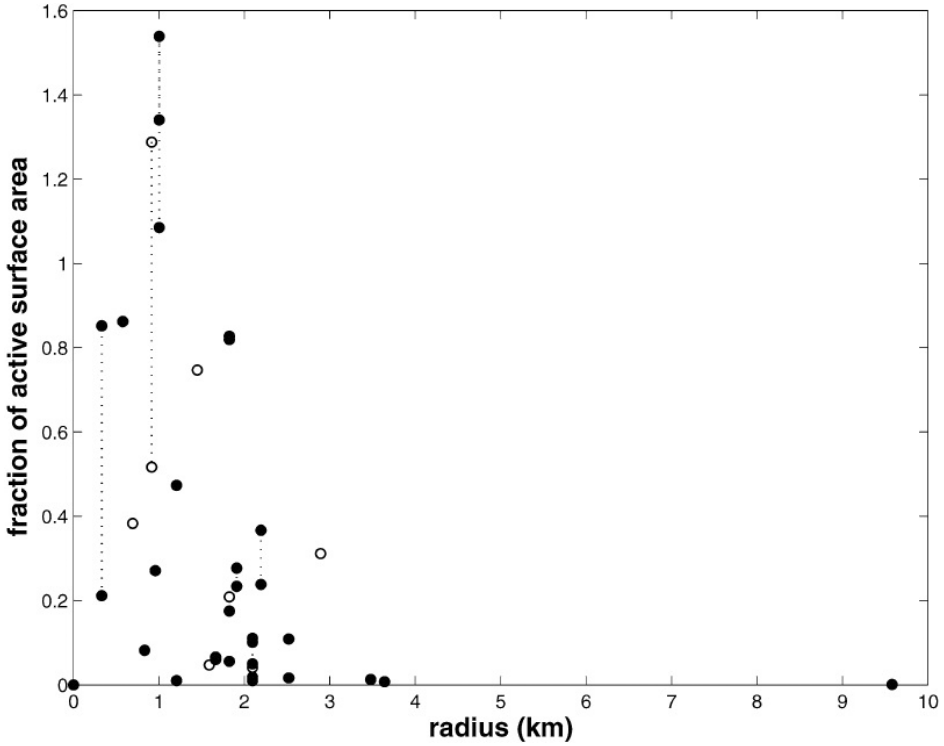


Fig. 2.7. Activities observed for Jupiter Family comets at $r < 2$ AU, using mainly the production rates from the near-UV OH band and the [O(^{1}D)] line according to A'Hearn *et al.* (1995). Open circles denote comets with nuclear radii of the lowest quality class, while filled circles are used for the other three classes. When more than one production rate is used, the individual symbols are joined by a vertical dashed line. The data point at the origin is erroneous and should be disregarded. Reprinted from Tancredi, G. *et al.*, *Icarus* **182**, 527–549 (2006), with permission from Elsevier.

comets that sometimes emit H_2O in quantities similar to what they would do, if their nuclei were icy spheres subject to sublimation due to solar heating.

Another comet of the same kind can be noted from recent observations including the EPOXI mission — comet 103P/Hartley 2 (see Sec. 1.5.4). While the physical mechanism behind the high water production rates cannot be clearly discerned for the three first mentioned comets, the situation is better in the last case, because it was established that much of the water came from sublimating icy grains in the inner coma (A'Hearn *et al.* 2011). In turn, as suggested by these authors, these grains might have been released from the nucleus by the very important CO_2 outgassing that the mission discovered. Of course, it cannot be excluded that other

Table 2.1. Nuclear parameters for comets that have been visited by space missions (sm) or have been subject to simultaneous visual and IR photometry (ph) in a state of very low activity, allowing accurate size determinations. The activities and albedos are global averages around the whole nuclei or their observed parts.

Comet	Nucleus dimensions (km)	Activity near perihelion	Geometric albedo	Source of information
1P/Halley	$15.3 \times 7.2 \times 7.2$	0.10	0.04	sm
9P/Tempel 1	7.9×4.2	0.03	0.06	sm
19P/Borrelly	8.0×3.15	0.3	0.03	sm
28P/Neujmin 1	19.2	0.001	0.03	ph
49P/Arend-Rigaux	9.2	0.007	0.04	ph
67P/Chur.-Geras.	4.3×4.1	0.06	0.06	sm
81P/Wild 2	$5.5 \times 4.0 \times 3.3$	0.25	0.03	sm
103P/Hartley 2	2.33×0.69	1.3	0.04	sm

Jupiter Family comets share the same behavior in the near-perihelion parts of their orbits.

In Table 2.1 we summarize the current estimates of physical properties for the best studied comets. These are not of homogeneous quality. Moreover, the thermal models used for individual comets are not the same. We will discuss comet 67P separately below, since this is by far the best studied case.

The most noteworthy information conveyed by these data is that the different comets have very different activities (like we saw in Fig. 2.7) but very similar albedos. The two quantities are not correlated, which verifies the impression that free surface sublimation is not the prime explanation of comet activity. Had it been so, the high-activity comets would also have had higher albedos due to the presence of ice on much of the surface, even if the albedo of the ice may be lowered by dust contamination.

The reason for the low albedo has not been fully elucidated. Since surface ice is rarely seen, we have two possibilities. Either the surface layers on all comets are made up of very dark material, or the light is trapped in the porous surface by multiple scattering together with absorption. Both effects are likely present. The organic refractories, forming an abundant constituent in comet material, may

well contribute importantly. The absence of the red spectral slope seen on small bodies that are considered to be related to comets, like Jupiter Trojans and D-type asteroids, may be explained by the freshness of the continually eroded comet surfaces, in case the other objects are reddened by long-term cosmic ray sputtering.

2.4.3. *The activity of comet 67P*

Dust jets in the inner coma of comets have frequently been used as indicators of local activity and sources of information about the spin of the nucleus. Such an analysis was carried out before Rosetta on comet 67P/Churyumov-Gerasimenko by Vincent *et al.* (2013) and led to the identification of three active regions at latitudes $+60°$, $\simeq 0°$ and $-45°$. These source regions were later localized more precisely by Lara *et al.* (2015) using Rosetta data. However, observing dust features is not the same as tracing gas sources on the nucleus surface.

The body of information on comet 67P from Rosetta is enormous. Activity was observed already during the approach phase in 2014. The MIRO instrument first detected water vapor on 6 June, when the comet was situated 3.92 AU from the Sun. The OSIRIS cameras observed that the nucleus was surrounded by dust already at the end of April and in July detected dust jets emanating from the nucleus at a distance of 3.7 AU from the Sun. Somewhat later, ROSINA detected H_2O, CO_2 and CO molecules in situ in the vicinity of the nucleus.

It soon became clear that the source of this activity was situated in the neck. Figure 2.8 (left panel) illustrates this by showing an early picture of the dust jets. Specifically, the Hapi region was identified. This is an area of smooth-appearing surface running along the neck with big boulders strewn along its central part — see the right panel of Fig. 2.8. It is close to the north pole and experienced a period of relatively high illumination at the time in question, though other areas on the nucleus were receiving even stronger insolation. In fact, the equinoxes of the 67P nucleus occur at $r = 1.7$ AU before perihelion and at $r = 2.6$ AU on the outbound branch. The northern hemisphere experiences a long summer, mostly in the outer parts of

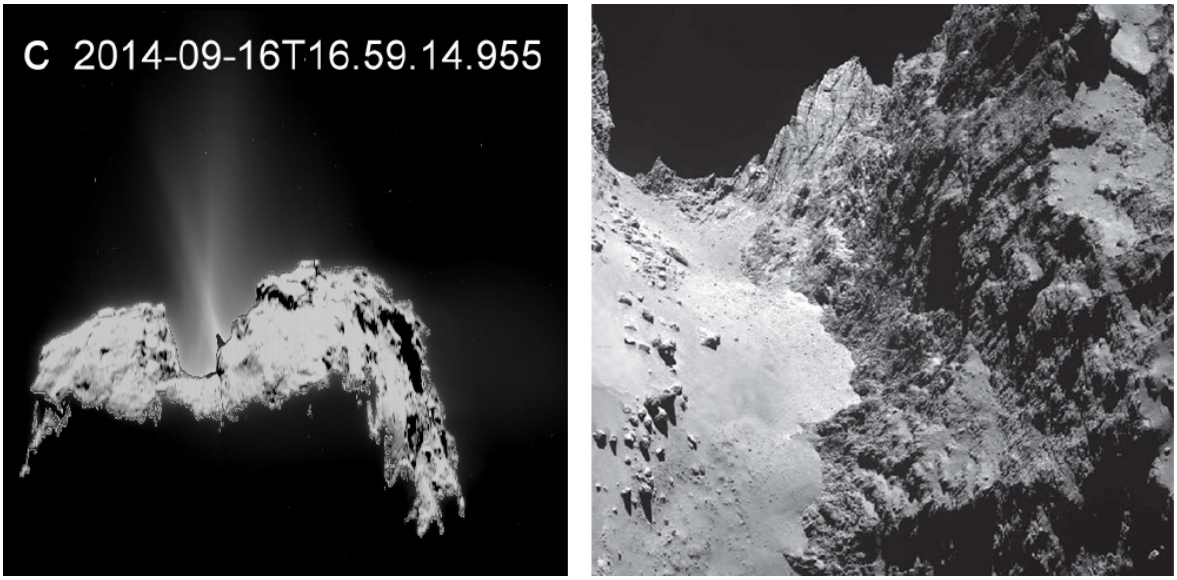


Fig. 2.8. Images of the 67P nucleus taken by the OSIRIS cameras. Left panel: Global view from 16 September 2014, showing a dust jet emanating from the neck region (wide angle camera). Credit: Z.-Y. Lin *et al.*, *A&A* **583**, A11 (2005), reproduced with permission © ESO. Extracted from Fig. 2 in Lin *et al.* (2005). Right panel: To the right, the Hathor wall, and to the left, the Hapi region (narrow angle camera). From Thomas, N. *et al.*, *Science* **347**, aaa0440 (2005). Reprinted with permission from AAAS.

the orbit, while the southern hemisphere is strongly insolated during perihelion passage and is therefore potentially more strongly eroded by sublimation.

A key to explaining the localization of the early activity is found in the shape of the nucleus. During August and September 2014, the Sun shone into the Hapi valley for part of each rotation, while the region was in shadow for most of the time. Thus, the integrated amount of direct insolation was rather low, but the surrounding walls of Hathor and Seth on the head and body, respectively, made an important contribution to the heating of Hapi by their thermal radiation and scattering of sunlight. This has been clarified by sophisticated thermal modeling, performed by Keller *et al.* (2015). The comprehensive model in question shows that the highest surface temperatures were mainly located in a band on the side of the Hapi valley closest to the Hathor wall under the assumption that the whole surface is ice covered. These temperatures are high enough to explain the activity by H_2O sublimation.

The fact that the activity of 67P started exactly where it should start assuming an isotropic surface, i.e., in the region of highest temperature, may indicate that the surface is indeed rather isotropic. This impression is supported by ROSINA measurements of gas density in the inner coma and their interpretation by gas dynamic simulations starting with outflow from the surface (Bieler *et al.* 2015; Fougere *et al.* 2016). From these it appears that the coma structure is best explained, if the sources of activity are widespread around the nucleus rather than restricted to a few active areas. Even so, the activity level of the surface was clearly non-uniform. Moreover, after the comet had approached the Sun close enough for the activity to become global (around New Year 2015), dust features were seen from all the illuminated parts of the northern hemisphere.

Meanwhile, the study of dust jets during the approach of the comet to perihelion (Vincent *et al.* 2016) has offered interesting, complementary information. With the exception of the broad, diffuse feature arising from Hapi, these jets could be resolved into very narrow structures, whose three-dimensional orientation could be discerned using series of images from the moving spacecraft. The footpoints of these jets were found to be concentrated to special topographic features, including nearly vertical cliffs or scarps. Roundish pits that are abundant in the Seth region form one category of features that jets arise from.

At least during the time in question, it appears that much of the distributed sources of outgassing were situated in such steep slopes — not only in the vicinity of Hapi. As a physical interpretation, Vincent *et al.* (2016) proposed a scenario, where flat areas are dust covered and therefore poor sources of outgassing, but cliffs expose fresh material. The latter would also be vulnerable to cracking by thermo-mechanical stresses, which facilitates the outgassing and also undermines the terrain behind the cliff. This may lead to collapse and mass wasting, whereby a layer of talus is deposited on the neighboring surface. Hence, as time proceeds, the topography levels out and the whole area may become dust mantled and thus inactive.

In the meantime, surface erosion proceeds in a lateral sense rather than the vertical erosion imagined by standard theories. This fits well

with results from other comets like 81P/Wild 2 and 9P/Tempel 1 (see Sec. 1.5). Formation of new pits, e.g., by sinkhole collapse (Vincent *et al.* 2015), may allow the process to continue and thereby save the comet from complete deactivation.

Whether the surface distribution of such local activity sources is actually uniform is a justified question. In particular, the gas production curve of comet 67P offers a chance to check. Keller *et al.* (2015) found that the uniform surface of their standard model does not produce a good fit to the observed curve. It fails to predict the amount of perihelion asymmetry and the fall-off slopes far from perihelion. Therefore, they considered a second thermal model that was computationally more efficient by sacrificing some details of the standard model but also more realistic from other points of view. This could be used in a search for an optimal distribution of activity over the surface to represent the observed H_2O production curve. Figure 2.9 shows that a reasonable fit to the data requires a spotted model instead of a uniform one. In the spotted models, a small number of directions from the nucleus center were selected as

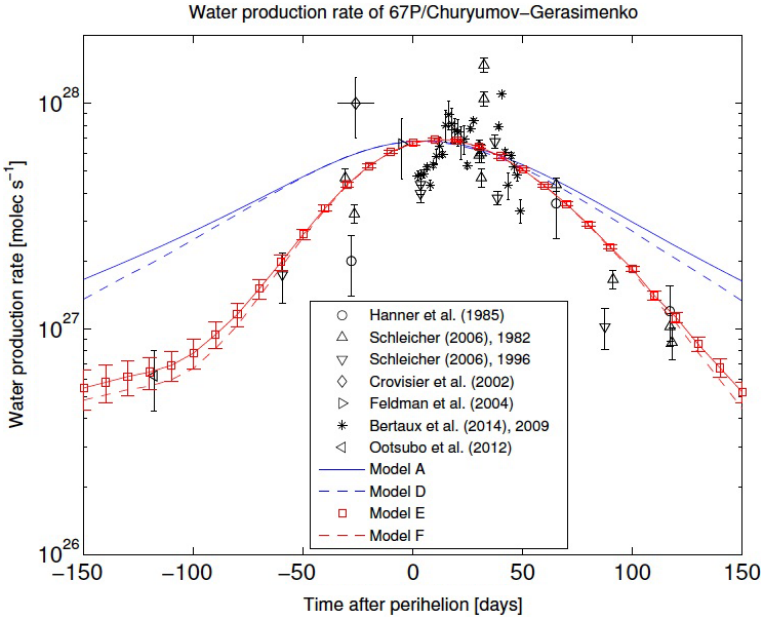


Fig. 2.9. H_2O production curve of comet 67P/Churyumov-Gerasimenko. The symbols indicate observations during preceding apparitions of the comet. The blue curves correspond to the model with a uniform surface, and the red curves illustrate the results of an optimal, spotted surface. Credit: H. U. Keller *et al.*, *A&A* **583**, A34 (2015), reproduced with permission © ESO.

centers of an active solid angle, while the rest of the surface had no activity.

Twenty-nine random surface activity maps were deemed successful in their fits to observations. From a combination of these it was seen that the local potential for activity has a maximum in a region joining the neck (opposite to Hapi) with an extended area on the large lobe. This coincides with the part of the nucleus, which is in permanently high illumination during the perihelion passage. Keller *et al.* (2015) found that there should be a physical explanation for this fact, and further discussion is given in Sec. 4.2.1.

The mentioned study of localized outgassing goes some way toward a complete activity model for comet 67P. Such a model should rather be unique and consistent with all the observations. However, the time may not yet be ripe for this endeavor. So far, detailed thermal models of cometary outgassing have not been coupled to hydrodynamic simulations of the coma, but this would be necessary for a full understanding of the activity whenever there is relevant information about the gas distribution in the inner coma.

The standard model of Keller *et al.* (2015) may not be ideal for determining the global activity parameter of comet 67P, since it does not yield a very good match to the H_2O production curve. However, it is the best as yet available and it was used for this very purpose. The result, based on gas production rates near perihelion, was about 6% — this is the value reported in Table 2.1. Alternatively, one may derive a value for the time integrated activity referring to the whole apparition as follows. The integrated mass loss of H_2O has been estimated as 2.7×10^9 kg (Bertaux 2015), while the standard model of Keller *et al.* (2015) with an icy surface produced an integrated loss amounting to 6.5×10^{10} kg. The ratio of these leads to a time integrated activity of about 4%.

Finally, we come to the issue of exposed ice on the surface of the nucleus. A very small amount of such ice was detected on the nucleus of comet 9P/Tempel 1 (Sec. 1.5.3), but the pre-perihelion search for H_2O ice by its infrared spectral signature by the VIRTIS instrument on board Rosetta only led to an upper limit of 1% of the area (Capaccioni *et al.* 2015). However, it takes only an extremely thin

layer of overlying dust to hide this signature, and thus, ice might be generally present very close to the surface, and small patches might even be ice-covered. Indeed, meter-sized icy patches were identified by OSIRIS as bright spots (Pommerol *et al.* 2015; Barucci *et al.* 2016) and by VIRTIS as small exposures of surface frost (DeSanctis *et al.* 2015; Filacchione *et al.* 2016a).

The situation changed, as the comet came to perihelion and the southern hemisphere was targeted by the Rosetta investigations. The surface of the nucleus gradually became more neutral (less red) in color on approach to perihelion and then again, took a redder hue on the way out (Fornasier *et al.* 2016) — a variation that reveals an increased presence of near-surface ice in the innermost part of the orbit. We note that a full thermal model aiming to reproduce the gas production curve of 67P should also incorporate this feature. In addition, two large ice-rich patches (about $1\,500\,\mathrm{m}^2$ each) were observed by OSIRIS shortly before perihelion (Fornasier *et al.* 2016). They lasted only for about 10 days, after which the ice had apparently sublimed. The patches were situated in the southern Anhur and Bes regions.

About three weeks earlier, VIRTIS had observed exposed CO_2 ice in one of these areas (Filacchione *et al.* 2016b) — again, only of short duration. The nature of this ice, whether it is a recondensed frost or a newly exposed patch of pristine material, remains to be explored.

2.5. Structure, Density and Porosity

A fundamental question about comet nuclei is how they came together. It is usually thought that this is part of the broad picture of planetesimal formation, which starts from micron-sized grains and leads to macroscopic objects the size of comets or even larger (see Chap. 7). The structural properties of comet nuclei may hold important information about this process. In particular, the way a typical nucleus is composed of sub-units of various sizes, the strength of the bonds that hold these together, and the porosity of the assemblage are items of prime interest. Let us now review the observational evidence that has a bearing on these.

2.5.1. *Morphologic evidence*

The first feature to strike the viewer of a picture showing the nucleus of comet 67P/Churyumov-Gerasimenko is that it looks like two nuclei glued to each other (Sec. 1.5.5). Such objects are called *contact binaries*. The elongated nuclei of 1P/Halley, 19P/Borrelly and 103P/Hartley 2 have all been put forward as candidates for a contact binary structure based on the close-up images, but only the last mentioned case can be deemed nearly as clearcut as that of comet 67P. Among other comets there is only one unquestionable case, namely, the nucleus of comet 8P/Tuttle. This showed its binary nature by radar imaging (Harmon *et al.* 2010) after Hubble Space Telescope observations by Lamy *et al.* (2008) had suggested such a nature.

It is important to note that the absence of evidence for contact binaries among the rest of the comets is no evidence of absence of this structure. Lacking close-up images, it is hard to conclude against a contact binary nature for any individual object. From the imaged cases, if one accepts that 9P/Tempel 1 and 81P/Wild 2 do not have contact binary nuclei, the extant small number statistics still indicates that contact binaries may be at least as common as monolithic nuclei. It is thus reasonable to include this feature among the constraints, when studying the formation of the nuclei.

Concerning other large-scale structural features on comet nuclei, we may note the layers observed on the 9P/Tempel 1 nucleus and those, possibly of similar nature, on the 67P nucleus (Massironi *et al.* 2015). These might be remnants of the accretion process, if the latter involved cometesimals that were flattened by the pressure exerted by the impacts, as in the "talps" model by Belton *et al.* (2007). On the other hand, evidence of non-flattened cometesimals is rare. One possible case on the 67P nucleus was mentioned by Davidsson *et al.* (2016). This consists of three convex surface features on the small lobe with diameters of a few hundred meters, which may be cometesimals that preserved their structural integrity to some extent during their accretion — see Fig. 2.10.

A small-scale feature of possible relevance seen on the Rosetta/OSIRIS images of 67P was named "goosebumps" by Sierks

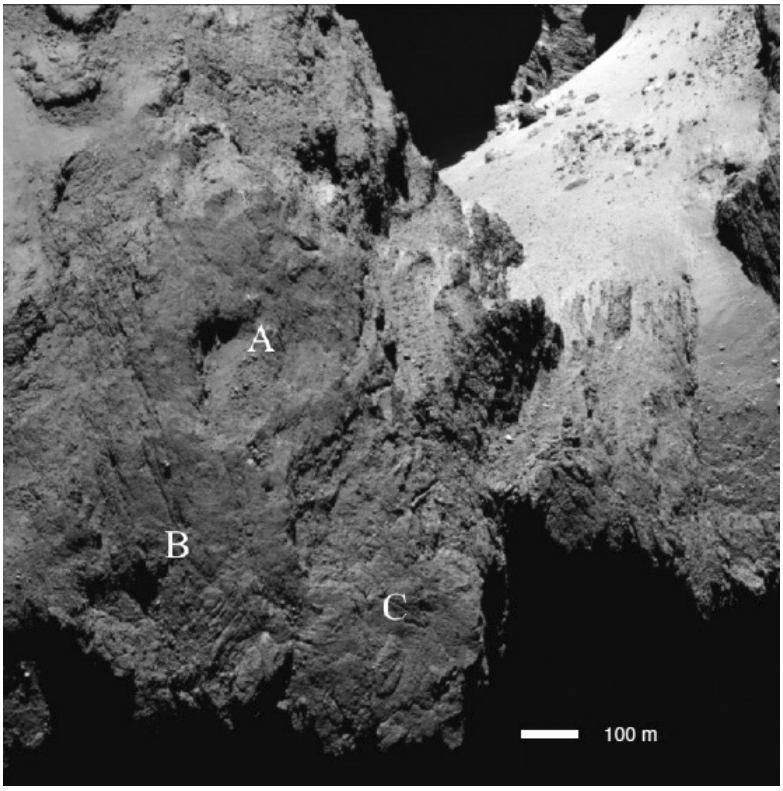


Fig. 2.10. Bastet region on the small lobe of the 67P nucleus (lower left), imaged by the OSIRIS narrow-angle camera. The letters denote three "positive relief features" that may be outcrops of individual cometesimals. The Hapi region is seen to the right. Credit: B. Davidsson *et al.*, *A&A* **592**, A63 (2016), reproduced with permission © ESO.

et al. (2015). An example is shown in Fig. 2.11. The typical size scale is a few meters, and they appear as more or less consolidated clods sitting side by side in large numbers but are only visible under favorable conditions like on the vertical walls of pits. There are several examples of goosebump fields in different regions of the nucleus (Davidsson *et al.* 2016), and the first question is how they were formed. Whether or not thermal cracking (El-Maarry *et al.* 2015) was involved, an intrinsic lumpiness of the comet material on the meter scale may be indicated with obvious consequences for understanding its origin.

2.5.2. *Strength and splitting*

When listing the phenomena that comets are known for, one cannot ignore the fact that they are prone to split apart. This has been known for 150 years, and recent time has witnessed several very

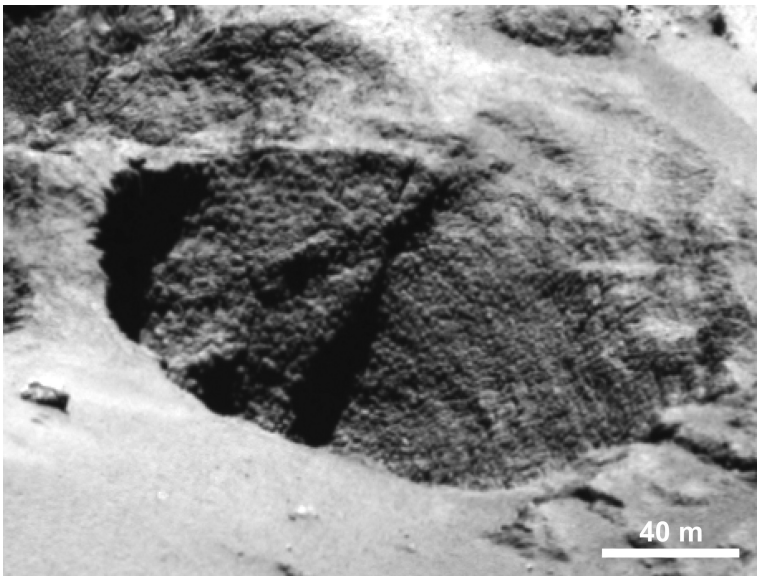


Fig. 2.11. Circular pit in the Seth region of the 67P nucleus, imaged by the OSIRIS narrow-angle camera. The grainy structure of the opposite wall is caused by meter-sized clods called goosebumps. From Sierks, H. *et al.*, *Science* **347**, aaa1044 (2015). Reprinted with permission from AAAS.

interesting observations of splitting events. Comet splits fall into two categories. In one case, they occur in the close vicinity of a massive, gravitating body (the Sun or a planet), and these are caused by the tension applied to the nucleus by the tidal force. In the other case, their occurrence seems random in space and time, and very little is known about the underlying reason. However, in both cases detailed observations and theoretical modeling have yielded information about the strength and buildup of the nuclei.

In March 1993 a strange comet, later to become one of the most renowned comets ever, was discovered in Arizona by Carolyn and Eugene Shoemaker together with David Levy. It is nowadays officially referred to as D/1993 F2 (Shoemaker-Levy 9). Upon discovery, the comet appeared like a string of about a dozen comets, which had to be physically related and thus signaled the recent split of a common parent. Soon enough, orbit determinations verified this hypothesis by showing that the comet was orbiting around Jupiter and had passed very close to the planetary surface in July 1992. This would allow the comet to be torn apart by Jupiter's tidal force.

The attention that comet Shoemaker-Levy 9 caught among the public and in media was mostly due to the fact that the following

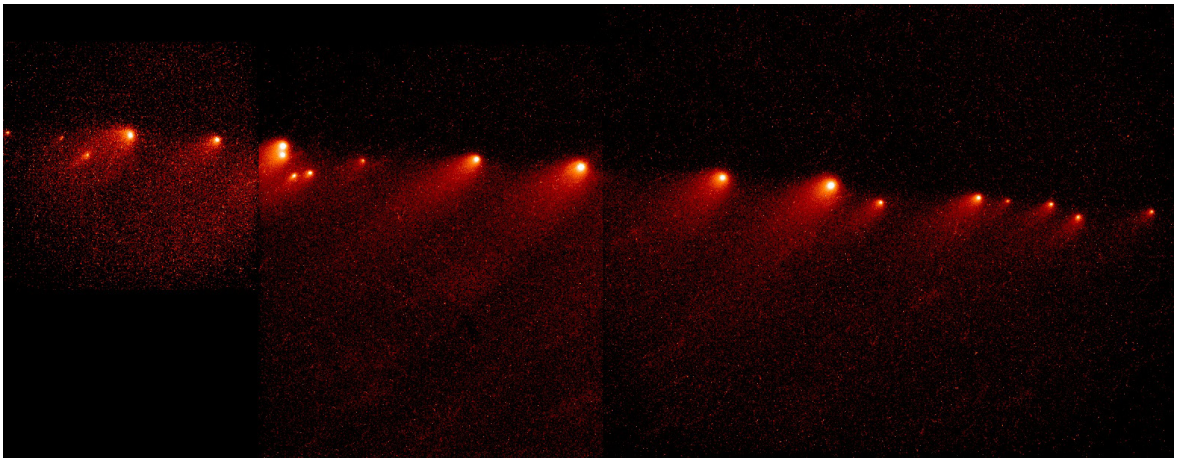


Fig. 2.12. A mosaic of images from the Hubble Space Telescope, showing the whole train of fragments that together make up comet D/1993 F2 (Shoemaker-Levy 9) on 17 May 1994, two months before these plunge into Jupiter's atmosphere. Photo Credit: H. A. Weaver and T. E. Smith (STSci), NASA.

perijove passage in July 1994 implied a series of collisions with Jupiter, which were accurately predicted and very well observed. The comet was shown to have orbited around the planet since several decades before the tidal breakup, and it provides one of the most prominent examples of temporary jovian satellite captures among observed comets (see Sec. 3.2.2).

A fundamental concept when discussing tidal splits is the *Roche limit*, which was introduced in the mid-19$^{\text{th}}$ century by French astronomer Édouard Roche. In the present context, the Roche limit is the smallest distance from the center of a planet, where a strengthless object can exist without being disrupted by the tidal force. Such an object is held together by its self-gravity, while the external gravity field in which it is placed tends to pull its surface away from its center along the line connecting the bodies. The latter is the tidal force, and it falls off rapidly (inverse cube dependence) with increasing distance from the planet. The distance at which the two forces are in equilibrium is the Roche limit (r_R). This is a small number of planet radii, but there is no universal value, since it depends on the density of the object. A standard expression is

$$r_R = 2.45 R_p \left(\frac{\rho_p}{\rho} \right)^{1/3}, \tag{2.13}$$

where R_p and ρ_p are the planet radius and density, respectively, and ρ is the density of the object.

If a comet approaches Jupiter very closely without ever passing inside the Roche limit, it will not undergo any tidal split. If it does penetrate within the Roche limit, it may survive intact due to its internal strength. If the strength is much smaller than the gravity forces and thus negligible, a split may or may not occur. Whether it does occur depends on detailed circumstances that may differ from case to case, but an important criterion is how far inside the Roche limit the perijove is situated.

There is a certain time around the perijove passage, during which the surface of the comet nucleus is accelerated away from the center in the two directions facing and opposing the planet. If this results in some material reaching the escape speed, the nucleus splits. After exiting through the Roche limit, the nucleus may reaccrete material moving at lower speeds, and concentrations of shattered material may fall together under self-gravity. According to Sridhar and Tremaine (1992), the shedding of material requires a perijove distance smaller than $0.69 r_R$.

The encounter of comet Shoemaker-Levy 9 with Jupiter in 1992 was the closest known for observed comets with a minimum distance of 1.3 jovian radii (20 000 km from the surface). This is deep inside the Roche limit for any reasonable density, so the split is no surprise as long as the strength is very low. It is of interest to compare with other comets, which are known to have made very close encounters with Jupiter. The only previous case of a tidal split is that of comet 16P/Brooks 2, which was discovered in 1889 as a multiple comet with several components. The background turned out to be an encounter with Jupiter in July 1886 with a perijove distance of 2.0 jovian radii. Sekanina and Yeomans (1985) modeled the separation of the components as a tidal split due to fracturing of a parent nucleus, following Aggarwal and Oberbeck (1974), and concluded that both the bulk density and the tensile strength of the Brooks 2 nucleus had to be very low.

For comet Shoemaker-Levy 9 the situation was much better, since much more observations were available to constrain the modeling of

the breakup. In this case Erik Asphaug and Willy Benz (1996) made the most comprehensive study, in which they first showed that the observed train of comets cannot be explained by any mechanism involving fracture in a monolithic precursor nucleus. They then applied hydrocode modeling of the tidal breakup of a strengthless precursor consisting of an aggregate of uniform, frictionless, spherical grains followed by the reaccumulation under self-gravity into a number of sizeable clumps. The outcome was studied as a function of the basic properties of the precursor: its size and density.

Assuming a non-rotating precursor, the best results were found for a diameter close to 1.5 km and a bulk density of $0.6 \pm 0.1\,\mathrm{g/cm^3}$. This would reproduce the basic characteristics of the train of comets: its length and the similar sizes of most of the members. With a higher density, a large central clump would form in contrast to the observations. However, a rapid prograde rotation of the precursor would offset this tendency, and a good fit could then be obtained with a 1 km diameter and a density of $1\,\mathrm{g/cm^3}$. Since the state of rotation is not constrained by any other observations, and the total mass of the precursor cannot be determined accurately enough from observations and modeling of the 1994 impacts, there is hence a range of possible densities from about 0.5 to $1\,\mathrm{g/cm^3}$.

In summary, the experience from comet Shoemaker-Levy 9 tells us that this comet had an essentially strengthless nucleus of low density. Its temporary jovian satellite capture does not suggest that its dynamical evolution could have bestowed special properties on it, different from most Jupiter Family comets. Therefore, one has good reason to believe that strengthless, low density nuclei are the rule, but of course, support for this picture from other comets is essential. In addition, care needs to be exerted concerning the term "strengthless". In the Asphaug–Benz model there is no strength whatsoever, but for real comets this only means that strength should be negligible compared to the already very small forces of the planetary tide and self-gravity.

The vast majority of comet splits are non-tidal. These occur at large distance from any massive object, covering a very wide range of heliocentric distance. Unfortunately, it has not been possible to reveal

any particular mechanism as the main culprit for these phenomena, and thus it is difficult to draw conclusions about physical properties based on their occurrence. However, one thing that is certain is that such random splits are quite common, possibly playing an important role in comet evolution. There are hence many cases of well-observed splits, from which valuable data can be derived. The most recent, general review was written by Boehnhardt (2004), who presented the relevant statistics and discussed their implications.

In particular, in many cases, the separation velocities, lifetimes, and accelerations of fragments relative to the primary component have been estimated. At first impression, there is no uniformity among these parameters — they all vary within a few orders of magnitude between different events. One thing that stands out (Boehnhardt 2004) is that the largest separation velocities (tens of m/s) are often associated with periodic comets. Among several splitting mechanisms that are often discussed, *rotational splitting* is perhaps the most popular due to its relative lack of problems explaining the observations. This involves a spin-up of the nucleus by a torque due to asymmetric outgassing until the centrifugal force exceeds the self-gravity at the equator, so that shedding of material may occur.

If there is virtually no cohesion in the nucleus, the shed material would leave at the equatorial spin velocity $V_{\rm eq}$, which in centrifugal equilibrium would be proportional to the nuclear radius R and amount to $0.36\,\mathrm{m/s}$ for $R = 1\,\mathrm{km}$ with a density of $0.5\,\mathrm{g/cm^3}$. Thus, without strength, the larger separation velocities would require unreasonably large nuclear radii. In fact, if the separation instead occurs by yielding of a cohesive material under the centrifugal force, the tensile strength can be estimated to reach the order of $10^5\,\mathrm{Pa}$ (Sekanina 1982; Boehnhardt 2004).

Even though this may appear a bit speculative, it is supported by the fact that the largest separation velocities are restricted to periodic comets. These are the ones that make a large enough number of orbital revolutions that the outgassing torque can build up a very fast spin. Long period and new comets would only be subject to rotational splitting, if they have very low strength so that a slower spin may do the job.

One may also obtain some information from the observed spin rates of comets concerning the minimum strength to hold them together. For the usual, km-sized nuclei the situation is rather inconclusive, since spin periods short enough to endanger their stability have not been observed. However, the very big nucleus of comet C/1995 O1 (Hale-Bopp) needs a strength similar to the above estimate to survive intact with its spin period of 11.5 hours (Boehnhardt 2004).

It is not entirely clear, what this means for the comparison with comet Shoemaker-Levy 9. One must keep in mind that the split of the latter involved the whole nucleus, while the rotational splits would only affect layers close to the surface. From the findings concerning comet 67P/Churyumov-Gerasimenko it appears that a surface layer of possibly considerable thickness is characterized by high-strength material, while the deep interior may well be strengthless.

2.5.3. *Density and porosity*

There is only one generally available method to determine the density of a comet nucleus, i.e., to determine its mass and volume and divide the two. However, in practice this approach is realistic only for a minority of the short-period comets, as far as individual densities are concerned. To see why this is so, we first consider the mass determination. Except in peculiar situations, this rests on observation and analysis of the nongravitational effects in the orbital motion together with the gas production curve. These are well known only for some of the short-period comets. To this comes the problem that the average radius of the nucleus has a large observational error bar for many of those, making volume estimates extremely uncertain.

Naturally, spacecraft targets are the preferential objects for density determinations. Not only do these present the most direct and accurate volume estimates, but their mass determination is also facilitated by the fact that, for preparation and follow up of the space missions, they are well observed both ground-based and from satellites. This is a guarantee for availability of good information about the gas production curve, which is necessary to interpret the nongravitational effects.

The basis for these mass determinations is Whipple's (1950) theory for the solid nucleus, in which the nongravitational effects observed in comets were interpreted as results of the jet force caused by asymmetric outgassing. To illustrate how this has been put into practice, we list some historical facts. The nongravitational effect that was first established was an offset of the time of perihelion passage from the ephemerides for returning periodic comets. In orbital solutions, this was first represented as an empirically determined correction to the mean motion that would make the orbital period longer or shorter as needed. By itself, this correction involved no physics.

The same can be said about the improved treatment introduced by Marsden *et al.* (1973). In Sec. 2.3.1 we discussed an extremely simplified thermal model for a comet nucleus, called isothermal, and in Fig. 2.3 we presented theoretical gas production curves based on this model for various sublimating molecules. As noted, the curve for H_2O is of particular interest, since it comes closest to portraying the observed activity variations of comets. An analytical representation of this curve, usually denoted $g(r)$, forms part of the Marsden *et al.* model.

The model implies that a nongravitational acceleration is introduced into the equations of motion for the comet. In comparison to the empirical corrections to the mean motion, this was a great step forward. However, there was no physical theory available to parametrize the expression for the nongravitational force, and hence an empirical approach was once again used. Here, the acceleration vector is described as $\mathbf{A} = (A_1, A_2, A_3)\, g(r)$, and A_1, A_2 and A_3 are constants representing the orthogonal radial, transverse, and normal components of the acceleration at $r = 1\,\mathrm{AU}$. These constants can be solved for along with the osculating orbital elements to fit the observed astrometric positions of the comet at various times. The influence on the orbital period means a perturbation of the orbital energy, and the A_2 term yields such a perturbation, while in a first approximation the influence of the A_1 term cancels out before and after perihelion, and the A_3 term has no influence at all.

That this model practically involves no physics can be seen from the fact that it is based on the isothermal model of sublimation

from the nucleus. In this approximation there is no asymmetry of the outgassing and hence no jet force. Moreover, there is no physical basis for assuming the constancy of the three parameters, and it can even be surmised that these would not be constant, if the acceleration vector was derived from a more realistic thermal model. Even so, the accuracy of the observations is limited, and as long as the fit is satisfactory, the model can be deemed satisfactory too. In fact, it quickly became the preferential model in almost all nongravitational orbit determinations and is usually called the *standard model*.

Calculations of non-uniform distributions of surface temperature and sublimation rate from a thermal model were first applied to the problem of the nongravitational force by Rickman and Froeschlé (1983b). Huge variations of A_2 as derived from the instantaneous force vector were found, but upon closer inspection it was concluded that the standard model cannot be replaced by an equally comprehensive, more realistic model (Froeschlé and Rickman 1986).

As mentioned above, the parameter A_2 is closely related to the perturbation ΔP of the orbital period during one revolution. Comets 1P/Halley and 22P/Kopff were the first, for which the mass determination was carried out (Rickman 1986). Halley's comet is known to arrive at perihelion about four days too late due to its nongravitational perturbation, so $\Delta P \simeq 4\,\mathrm{days}$. Using the Gauss equations (McCuskey 1963), the following formula can be derived for calculating the nongravitational perturbation of the orbital period from the radial (F_r) and transverse (F_t) components of the jet force in the orbital plane:

$$\Delta P = \frac{6\pi\sqrt{1-e^2}}{Mn^2} \cdot \left\{ \frac{e}{p} \int_0^P \left(F_r \sin f + \frac{F_t}{r} \right) dt \right\}, \tag{2.14}$$

where M is the mass of the nucleus, t is time, e, n and p are the eccentricity, mean motion and semilatus rectum of the comet orbit, f is the true anomaly, and r is the distance from the Sun.

The jet force can formally be expressed as

$$\mathbf{F} = -Q\, m\, \mathbf{u}, \tag{2.15}$$

where Q is the outflow rate of gas molecules from the nucleus, m is the mean molecular mass, and $\mathbf{u}$ is the effective outflow velocity, which means the vectorial average of all the velocities of individual molecules as they leave the nucleus. This definition was given by Rickman (1989), but a few comments are in order. It is not obvious that Q can be identified with the observed gas production rate of a comet, because some of this gas may have been produced in the coma by decomposition or sublimation of solid grains. The averaging of the velocities is a complicated procedure, since active comets near the Sun have a so-called *Knudsen layer* next to the surface, where the individual molecules get together into the bulk outflow of an equilibrium gas by collisional mixing. This involves a back flow to the nucleus, which adds to the momentum exchange. Moreover, if Q refers to the observed H_2O production, there may be other, unobserved gases that also contribute to the jet force.

The mass M can be obtained from Eq. (2.14), if F_r and F_t are derived from a combination of observations and thermal modeling of the nucleus. In Eq. (2.15), the gas production curve $Q(t)$ essentially comes from observations, while u_r and u_t come from modeling. For comet Halley, it turned out that the accuracy of the mass determination is mainly limited by the lack of consistency between different data sets concerning the H_2O production rate (see Sec. 4.1). Depending on which data is preferred, the analysis by Rickman (1989) gave masses from about 1 to 3 times 10^{14} kg.

The volume of the 1P/Halley nucleus was determined from the *in situ* imaging with thus far unprecedented accuracy as 365 km^3 (Merényi *et al.* 1990), and thus the bulk density of the nucleus could be crudely estimated as 0.3–0.7 g/cm^3. An update was made by Skorov and Rickman (1999), using new modeling of the Knudsen layer, whereby the density range was raised to 0.5–1.2 g/cm^3.

In the 21st century, new thermal models of comet surfaces were developed by Björn Davidsson, and within this framework new mass determinations for spacecraft targets were made by Davidsson and Pedro Gutiérrez in a series of papers (2004, 2005, 2006). Here the modeling was made self-consistent in that activity maps of the nuclei were made to fit the observational data on H_2O production rates

and spin axis orientations as well as to predict the nongravitational effects (not only the perihelion delay) for comparison with the available orbit determinations. For 19P/Borrelly the density came out as 0.2–0.3 g/cm^3, for 67P/Churyumov-Gerasimenko they found < 0.5–0.6 g/cm^3, and for 81P/Wild 2 the result was < 0.6–0.8 g/cm^3. In a following paper, Davidsson *et al.* (2007) found 0.2–0.7 g/cm^3 for 9P/Tempel 1.

There is a fairly good consistency within this set of density determinations, although there is of course some model dependency due to the assumptions used for the nature of cometary outgassing. However, there is also independent support from a few other cases. One of these is comet Shoemaker-Levy 9, which was discussed above. For comet Tempel 1, the mass of the nucleus was estimated from the Deep Impact observations of the expansion of the base of the ejecta plume as 0.4 g/cm^3 with an uncertainty range of 0.2–1.0 g/cm^3 (Richardson *et al.* 2007), and the new volume estimate from Stardust-NExT led to a slight revision upward (Thomas *et al.* 2013). The most solid result so far has recently been obtained for comet Churyumov-Gerasimenko from the mapping of the gravity field by the Radio Science Investigation and the OSIRIS imaging, and the value is 0.533 ± 0.006 g/cm^3 (Pätzold *et al.* 2016).

Hence, it now seems well established that comet nuclei have low densities, but of course, if the density should be called low depends on what material the nuclei are made of. In fact, the porosity is the interesting quantity, meaning the fraction of the volume that is not occupied by solid material (gas can be present in the pores, if the temperature allows the sublimation of volatiles from the pore walls). Assessing the porosity of comet nuclei involves considering two critical issues that are not yet fully understood. The first is the question just raised: which substances contribute to the comet material, and in which proportions? Answering this question means providing a value for the compact density of the solids in the nucleus, and the porosity is then the ratio between the actual bulk density and this compact density.

The second issue deals with the extent of the voids causing the porosity. It may be a question of pores on the microscale, like those

occurring when micron-sized grains are packed in a loose structure. But there may also be macroscopic voids, if the nucleus is a rubble pile where large cometesimals sit together without much compaction. Finally, one needs to distinguish the bulk density from the local density of surface units or the average bulk density of the surface layer. The latter can possibly be both larger and smaller than the bulk density of the whole nucleus. Surface layers can be compacted by the effects of cooking or sintering, at least locally, and this may contribute to the elevated range of surface densities (~ 0.5–$1.5\,\mathrm{g/cm^3}$) found from radar observations by Harmon *et al.* (2004).

Concerning the issue of the abundance of refractories versus ices in the comet material, the Rosetta exploration has brought some preliminary insight. The dust/gas ratio in the outflowing material of comet Churyumov-Gerasimenko before perihelion was reported by Rotundi *et al.* (2015) as 4 ± 2. Using their accurate bulk density determination, Pätzold *et al.* (2016) found, assuming the measured dust/ice ratio to hold throughout the nucleus, that the porosity of the comet would be 72–74%. The material would be quite dusty, since the inferred dust/ice ratio by volume is about 2.

Concerning the large-scale structural homogeneity of the nucleus, Pätzold *et al.* (2016) found that the C_{20} and C_{22} coefficients of the comet's gravity field agreed quite well with those inferred from the shape model assuming a constant density. Therefore, it seemed that any internal voids large enough to cause departures in these coefficients could be excluded, and this places significant limits to the amount of macroporosity in the nucleus. The CONSERT measurements (Kofman *et al.* 2015) are in agreement with both the homogeneity and the dust/ice ratio in the upper layers of the minor lobe of the nucleus.

2.6. Chemical Nature of Comets

The material of comet nuclei can be divided into three categories: ice, organics and rock minerals. The ice is fundamental to the picture of the nuclei as introduced by Whipple (1950), and the other two contribute to the solid particles observed in the comae and tails.

Silicate minerals have been taken for granted since long ago and supported by spectroscopic observations of Sun-grazing comets, while the organics were discovered more recently. They can be analyzed one by one in terms of chemical and mineralogical composition.

2.6.1. *Cometary volatiles*

Being the most abundant, H_2O is the natural reference molecule for cometary volatiles. But the abundances of other volatiles relative to H_2O in the ice cannot be directly observed. In the best case, the observations refer to simultaneous production rates from the nucleus, but even so it is not obvious that different molecules are produced by the same mechanism at the same place. This would be the case, if the other molecules existed in the nucleus as guest molecules in clathrate hydrates formed of hexagonal water ice. Sublimation of the ice at the surface would then yield the other molecules in the same proportions as they occupy as guests in the ice. However, the truth is not so simple.

The Christmas tree diagram in Fig. 2.5 brings out this fact very clearly. The production curve of H_2O (traced by OH) is not parallel to the other curves in comet Hale-Bopp. For CO one question is evident: Is this more abundant or less abundant than H_2O? To answer this question, we may consider that the integrated productions of all molecules during the whole apparition correspond to the erosion of one and the same surface layer, so that the ratios of different integrals reveal the abundance ratios in that layer. However, the assumption made is not necessarily correct. Alternatively, we should have a good idea about the actual processes behind the release of the different molecules, allowing to estimate their efficiencies quantitatively.

Neither of these alternatives is generally available, however. Therefore, the best remaining option is to simply compare the observed production rates in comets close enough to the Sun for all molecules to be released into space. Such a compilation, made by Dominique Bockelée-Morvan and co-authors (2004), is shown in Fig. 2.13. We note the high abundance of oxidized carbon (CO and CO_2) relative to the reduced (hydrogenated) species. The molecules

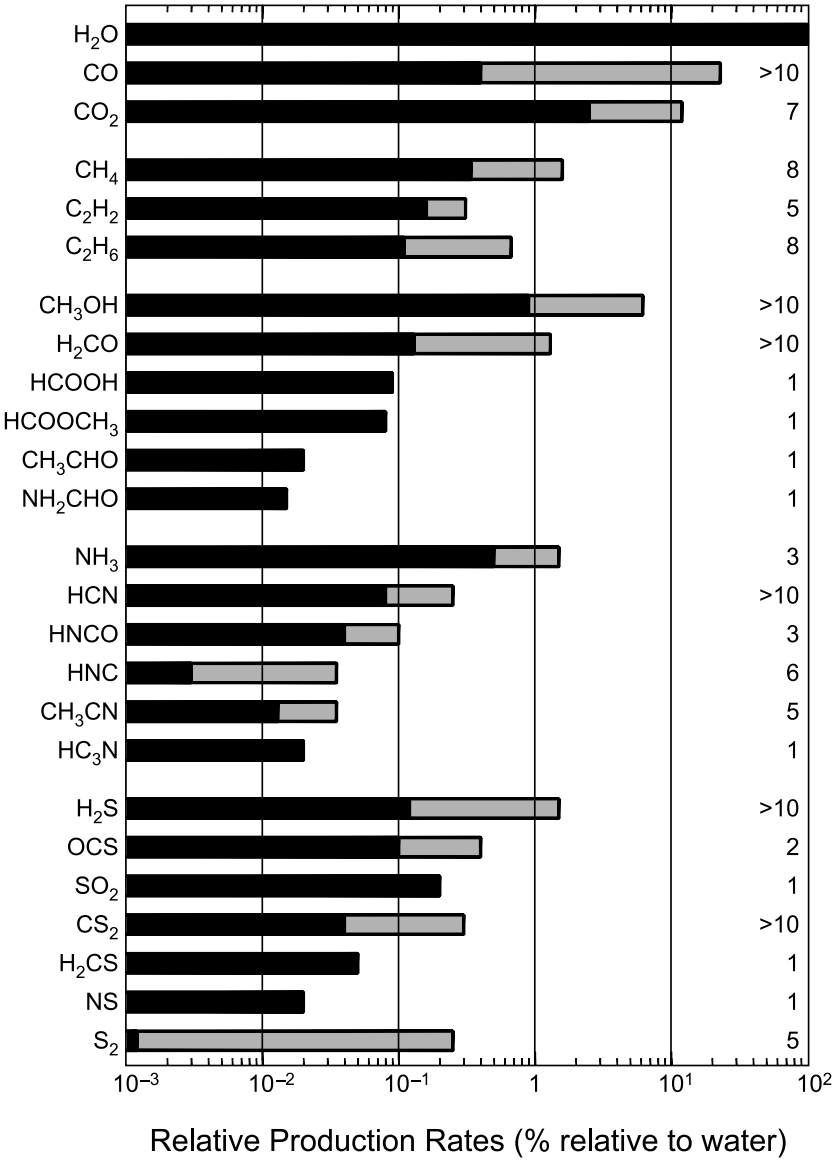


Fig. 2.13. Observed abundances of molecules relative to H_2O as found from spectral detections in various comets. The number of individual comets used for each specific molecule is shown at the right edge of the panel. From D. Bockelée-Morvan *et al.*, "The Composition of Cometary Volatiles," in *Comets II*, edited by M. C. Festou *et al.* © 2004 The Arizona Board of Regents. Reprinted by permission of the University of Arizona Press.

are generally outgassed from the nuclei, but for CO and H_2CO there is often an important distributed source in the form of organics, and HNC is produced by coma chemistry.

Historically, CO_2 observations have in general been made more recently than those of CO. Of special importance was the Japanese *AKARI* satellite (Ootsubo *et al.* 2012). An analysis of CO and CO_2 abundances in all types of comets relative to H_2O was presented by

Mike A'Hearn and co-workers (2012). This showed a large scatter of individual values for both molecules as well as of the ratios between CO and CO_2 in the same comet. No correlation between the two is evident, and no clearcut difference between Jupiter Family comets and other types is found.

Both CO and CO_2 were observed along with H_2O in comet 67P/Churyumov-Gerasimenko by the ROSINA mass spectrometer on board the Rosetta orbiter. These three molecules were clearly identified as the dominant ones in the cometary coma during the early part of the investigation. The instrument observed gases coming from the nadir direction on the nucleus, which changed with latitude and longitude due to the orbital motion of Rosetta and the spin of the nucleus. Strong temporal variations of the relative detection rates of both CO and CO_2 with respect to H_2O were observed (Hässig *et al.* 2015), while both molecules were on the average very important in this remote part of the cometary orbit. In particular, a pronounced dichotomy of the nucleus was identified, whereby the outgassing from the northern hemisphere (specifically, the Hapi region) was H_2O dominated, while the other molecules predominated near the unilluminated south pole. Is this a sign of chemical heterogeneity in the 67P nucleus? This question will be discussed in Sec. 7.4.2.

When several comets are available for comparison, the abundances often differ substantially. The thermophysical properties of the different molecules span a wide range of saturation pressure, or volatility. An extreme case is the S_2 molecule, which is highly variable but reached 0.25% of water in comet C/1983 H1 (IRAS-Araki-Alcock) according to Budzien and Feldman (1992). This molecule has a very low equilibrium condensation temperature. Comparing carbon and nitrogen is of interest. Their hydrides, methane and ammonia, are often detected at about 1% of H_2O, but while pure carbon in the form of graphite is common in the refractory component, molecular nitrogen — like molecular oxygen — took a long time to be detected in any comet.

This happened with the *in situ* Rosetta exploration of 67P, specifically by ROSINA. Both molecules were detected but at quite different abundances. While O_2 was found to be produced at a ratio

of 3.8% with respect to H_2O (Bieler *et al.* 2015) and was subsequently found to have been produced at 3.7% also in comet 1P/Halley (Rubin *et al.* 2015a), the production rate of N_2 in comet 67P was only ∼1% with respect to CO (Rubin *et al.* 2015b). Moreover, argon was for the first time identified in a cometary coma. Its production rate was ∼1% with respect to N_2 (Balsiger *et al.* 2015). How to best interpret these results will be discussed in Sec. 7.3.2.

2.6.2. *Cometary silicates*

A real insight into the nature of the cometary silicates had to await the invention of relevant spectral IR-observing facilities, in particular the Hawaii-based NASA Infrared Telescope Facility and the *Infrared Space Observatory* (ISO). Fortunately, several bright comets including Hale-Bopp were thus observed with good spectral resolution in the region of silicate emission around 10 μm. Later on, equally important observations have been made by the *Spitzer Space Telescope*, and the *Stardust* sample return mission brought fundamental information on the composition of the most refractory coma grains in comet 81P/Wild 2.

From early IR data, specifically 7.8–13 μm spectra of seven comets, Hanner *et al.* (1994) concluded that amorphous silicates of both olivine and pyroxene type were very common, but that three comets also showed features indicative of crystalline olivine. The former type dominates in interstellar space and is likely the native form of the silicates. The latter type, on the other hand, requires high temperature processing — for instance, the evaporation and subsequent recondensation of the silicates according to the common scenario for equilibrium models of solar nebula chemistry. However, such processes would only occur close to the Sun, while comets should have formed much further out.

The presence of both amorphous and crystalline silicates was confirmed by spectra of comet Hale-Bopp (Wooden *et al.* 1999). Finally, the Stardust samples from comet 81P/Wild 2 showed that the grains emitted by this comet contained both amorphous and crystalline silicates (Zolensky *et al.* 2006) — the latter very often in

the form of almost pure magnesium olivine called forsterite, which is a high-temperature condensate. One of the tracks formed in the silica aerogel was especially remarkable, since the responsible particle seems to have had a composition similar to the extremely refractory, meteoritic *Calcium–Aluminum-rich Inclusions* (usually called CAIs).

Of the low-temperature hydrated silicates that in meteorites appear as products of *acqueous alteration*, none were seen. This simply means that the comet material never saw liquid water, which is not surprising. But the intimate mix of unaltered silicates with variants formed at very high temperatures is a fascinating discovery. The altered material appears to have been transported from the place near the Sun, where it was formed or the alteration took place, all the way into the cold regions where the comet nucleus was aggregated. We shall return to this in Sec. 7.3.2.

2.6.3. *Cometary organics*

Each of the Giotto and Vega missions to comet Halley carried on board a time-of-flight ion mass spectrometer for roughly micron-sized dust grains. The results mentioned in Sec. 1.5.1 are illustrated in Fig. 2.14, from Kissel *et al.* (1986). Three single grains are shown in terms of the mass distributions of their atoms. They represent the three basic kinds of grains that were identified. One kind is obviously made of silicate minerals, being dominated by the rock-forming elements Si, Fe and Mg. Another type features the light elements H, C, O and N, and these are the *CHON grains*. The third kind is a mixture of the two components. The mixed kind was the most abundant followed by the CHON grains, while the pure silicate grains were rare.

Just like the silicate grain atomic compositions cannot be uniquely synthesized into specific minerals, the CHON grains can only vaguely be referred to as "organics". Furthermore, having survived expulsion from the nucleus in the solid state, they are obviously more or less refractory, so more precisely they have been called *organic refractories*. Their exact nature could not be established from the Giotto data. Since they mostly are not refractory enough to have survived the entries into the Stardust aerogel (Sec. 1.5.2),

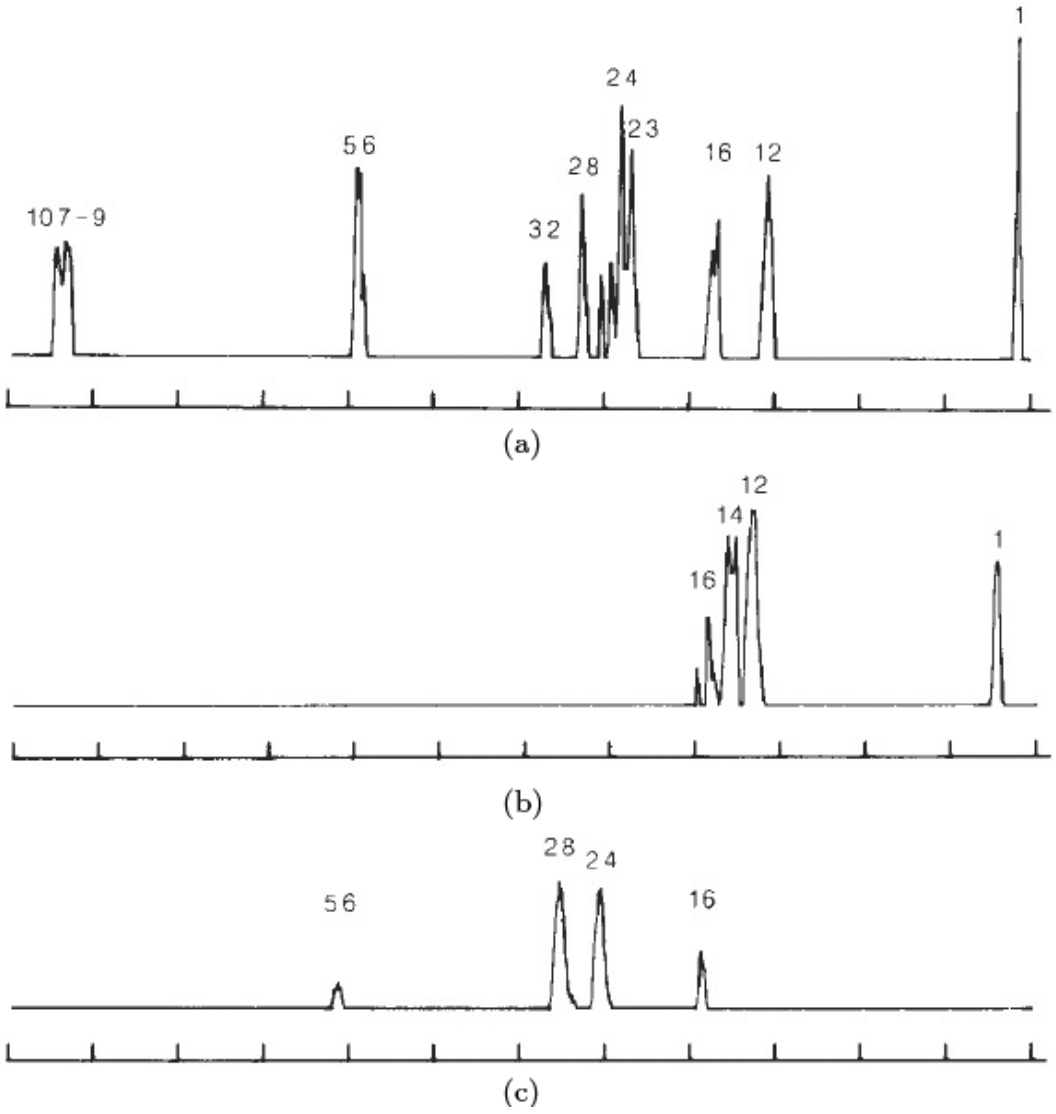


Fig. 2.14. Time-of-flight atomic mass spectra of three grains in comet Halley recorded by the PIA spectrometer on board Giotto. Panels *a*, *b* and *c* show mixed, CHON and silicate grains, respectively. The intensity scales are logarithmic, and the 107–109 mass peak for grain *a* is caused by silver ions from the target plate. Adapted by permission from MacMillan Publishers Ltd: Kissel, J. *et al.*, *Nature* **321**, 336–337 (1986), © 1986 Nature Publishing Group.

the analysis of returned Wild 2 material yielded only sparse and biased information (Sandford *et al.* 2006). However, the presence of nitrogen-rich polycyclic aromatic hydrocarbons (PAHs) was established along with a component that is poor in aromatics. Amino acids including glycine were also reported as found in Stardust-returned foil samples that captured gaseous material in the comet coma (Elsila *et al.* 2009). Overall, the organic material captured in the Stardust aerogel was found to be more akin to that of interplanetary dust

particles of cometary origin than to the organic residues derived from carbonaceous chondrite meteorites (see below).

Some more evidence came from spectral imaging of the coma in Halley's comet. It was found that the light emission from some radicals that contribute to the visible gas coma (like C_2, C_3 and CN) was anisotropically distributed around the nucleus. There was a good correlation between the maximum radical brightness and the dust jets seen in continuum images. This shows that the parent molecules of those radicals are not only emitted from the nucleus but also released from dust grains in the coma. However, the specific parent molecules and their abundances cannot be established with certainty.

Near-infrared spectroscopy of the 67P nucleus was performed within the Rosetta mission by the VIRTIS instrument (Capaccioni *et al.* 2015). From the albedo and spectral slopes seen across more or less the whole surface, it can be concluded that a pristine, organic refractory material is seen, which is different from tholins or other products of radiative processing. According to Quirico *et al.* (2016), it is a dark refractory, polyaromatic, carbonaceous material mixed with opaque minerals like iron sulfides and iron-nickel alloys. As illustrated by Fig. 2.15, between 2.9 and 3.6 µm wavelength there is an evident

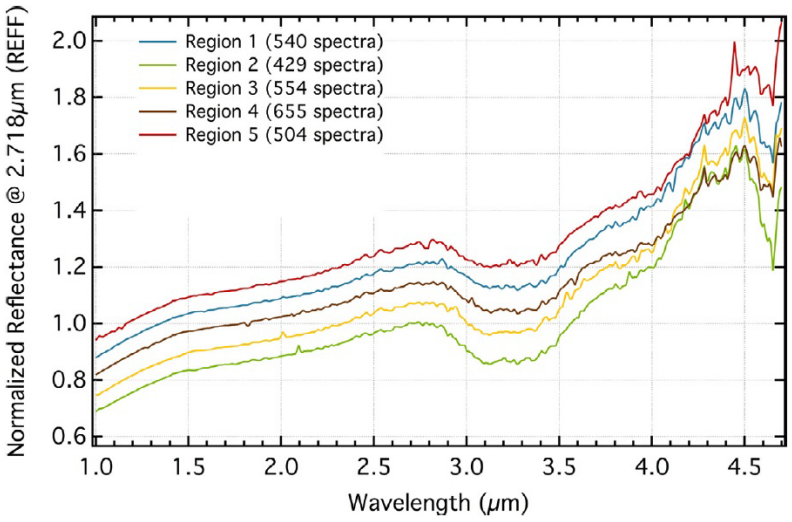


Fig. 2.15. Vertically shifted, near-infrared mean spectra of five regions on the 67P nucleus obtained by the VIRTIS instrument on board Rosetta. Reprinted from Quirico, E. *et al.*, *Icarus* **272**, 32–47 (2016), with permission from Elsevier.

absorption band, centered at 3.2 μm, which previously has not been seen on comet nuclei.

The paper by Quirico *et al.* provides an in-depth discussion of the likely carriers of this absorption. Like in the Stardust analyses, this is based on comparison with analog materials of known composition. These include both substances made in the laboratory and cosmic samples. Among the latter, there are the so-called *insoluble organic matter* (IOM) samples derived from carbonaceous chondrites, and the *interplanetary dust particles* (IDP) collected by stratospheric aircraft and *Antarctic micrometeorites* (AMM), which certainly contain an important fraction of cometary material. In the case of the 3.2 μm band, none of these analogs provides a satisfactory match, and preference is given to a semi-volatile component, plausibly containing carboxylic acids and the NH_4^+ ion. Interestingly, no hydrated minerals have been identified, and no genetic link with the CI, CR and CM chondrite types has been established, consistent with the above-mentioned Stardust results. There is no evident ice absorption in the spectra obtained early in the mission, but mapping of the 3.2 μm band across the nucleus surface indicated that water ice was a significant contributor in the neck region though more or less absent elsewhere.

The semi-volatile component of cometary organics is apparently of great importance. Its decomposition in the solar heat may release radicals like CN, C_2 and C_3, which tend to appear in comets within a few AU of the Sun. The case for this is the above-mentioned association of those radicals with dust features in the coma. If the semi-volatiles thus contain unsaturated carbon-chain molecules hosting C_2 and C_3, one may suspect the presence of a preserved interstellar component of comet material (Mumma and Charnley 2011). Moreover, below the surface of the comet nucleus, the organic semi-volatiles may act as a sintering agent, being decomposed at the very surface and providing material that to some extent diffuses downward through the porous matrix before condensing at grain interfaces.

One puzzling result came from a photometric narrow-band survey of production rates of the mentioned radicals and others

in many comets by A'Hearn *et al.* (1995), and largely confirmed by the spectroscopic survey of Fink (2009). These indicated that comets belong to either of two different categories. In one case, the production rates $Q(C_2)$ and $Q(C_3)$ vary proportionally to $Q(CN)$ in a typical way, whereas in the other case, these two radicals are clearly depleted relative to CN. It was found that essentially all long-period and Halley Type comets behave in the typical way, but there are two kinds of Jupiter Family comets: typical and depleted. This is illustrated by Fig. 2.16. All the radicals are largely produced from organic grains, but in some comets these grains are depleted in carbon chain molecules. Those comets strongly prefer the Jupiter Family in the sample analyzed by A'Hearn *et al.* (1995) but less so in a more recent sample (A'Hearn, private communicaton).

2.6.4. *Cometary material: The interstellar model*

There are two competing scenarios for the origin of the material that comet nuclei are made of. Let us first deal with the one

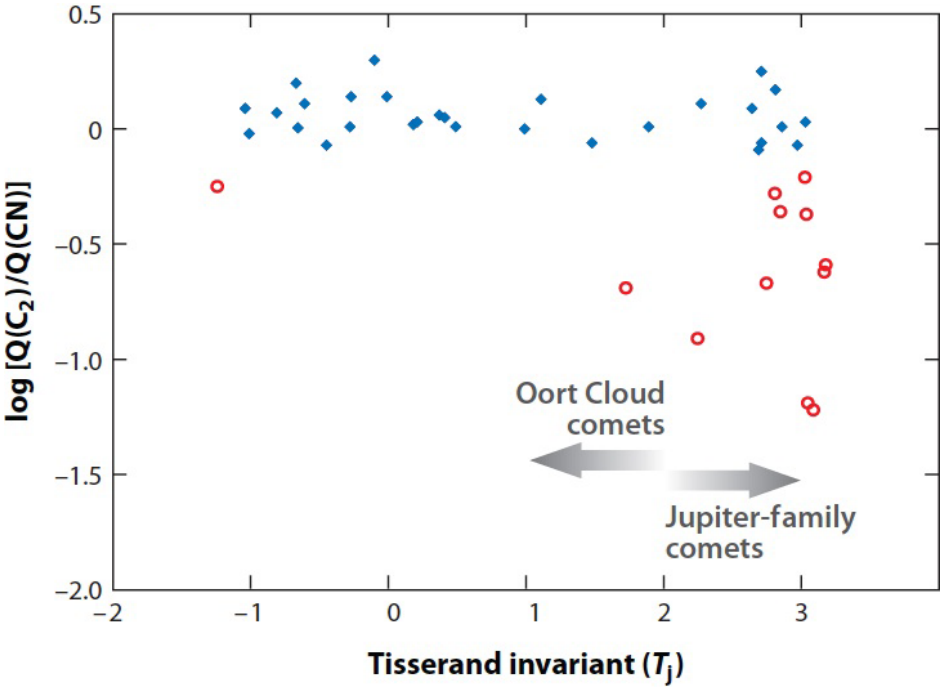


Fig. 2.16. Relative production rates of C_2 and CN in comets observed by A'Hearn *et al.* (1995), plotted versus the jovian Tisserand parameter. Carbon chain depleted comets are marked by red open circles. Credit: M. J. Mumma. Reproduced with permission of *Annual Review of Astronomy and Astrophysics*, Vol. 49. © by Annual Reviews.

that was favored in the wake of the Halley exploration, mainly because of the similarity between the compositions of comets and the prestellar material observed in star-forming regions. Its popularity has decreased, but it is still a strong contender to offer at least part of the explanation. We may call it the *interstellar model*.

The story begins in the extended, cool atmospheres of late-type giant or supergiant stars. This is an environment, where tiny grains of silicates or graphite may condense, and these get trapped in the outflow (stellar wind) that observations have revealed. Hence, they reach into interstellar space as members of the general population of interstellar grains. Such grains are known to exist since almost one century, because they cause an extinction of the light from distant stars, and this extinction is wavelength dependent, so that the stars get reddened. From the amount of interstellar reddening, a typical size of $\sim 0.3\,\mu$m has been estimated for the grains.

The properties of the interstellar gas vary from region to region, and this concerns in particular the density. But the grain/gas mass ratio remains about the same at $\sim 1\%$. The most important regions for grain evolution are the *dark molecular clouds*. These are the largest and densest of all, and from a vantage point inside such a cloud the grain population is effectively like an opaque wall preventing all visual and ultraviolet light of stars from entering. This means that there is very little heating of the gas, which takes an equilibrium temperature not far from absolute zero (specifically, $\sim 20\,$K).

With a relatively high gas density, an extremely low temperature and an abundance of grain surfaces, the conditions are excellent for forming molecules — in the first place, simple molecules like H_2 or OH. Due to the absence of UV radiation, such molecules may survive for a considerable time and are thus able to move around in the gas phase and meet other molecules. Thus, chemical reactions may occur, and new molecules may be formed. The high gas density implies that such reactions can proceed at an important rate, thus changing the chemical composition of the gas.

In fact, hundreds of molecules have been observed in interstellar gas, in particular in protostellar regions within giant molecular

clouds. Thus, more and more has been learned about the relevant chemistry, which is very different from laboratory chemistry. The latter proceeds in equilibrium due to the furious rate of interactions at high densities, but the molecular clouds present conditions that are truly vacuum-like. This means that extremely reactive radicals can exist for some time in the absence of collisions with other species, so these can participate in the chemical reaction network in a way that is very different from equilibrium chemistry. Another consequence is the survival of species like CO, which is found in large abundance in spite of the ubiquitous presence of H_2, which in equilibrium would transform this into CH_4 and H_2O.

The gas molecules are bound to collide with the grains. They often stick to the grain surfaces and thus change the composition of the grains by forming *icy mantles* on the pre-existing *refractory cores*. Due to the extremely low temperature, the rate of sublimation from the grain mantles may be very low even for very volatile species. Hence, sticking of new molecules may proceed faster than sublimation, and the mantles grow in thickness. Spectroscopic observations of proto-stellar objects have revealed the presence of several molecules in both gaseous and solid phases.

Thus, over long enough time, interstellar grains may become incorporated into molecular clouds over and over, and this provides for an important evolution. During a visit into a molecular cloud an icy mantle forms on the grain as seen, and this will include extremely reactive radicals. At the extremely low cloud temperatures these radicals stay practically immobile and thus do not react. However, they retain a large chemical potential, should the grain be heated. Now, the molecular clouds do not live forever. To some extent, they are consumed by star formation, and they tend to be dispersed by the shock waves around supernovae, which arise quite rapidly due to the short lifetimes of the most massive stars. When a molecular cloud is thus dispersed, lots of ice-mantled grains find themselves in a new environment with much less extinction, and the temperature increases.

As a consequence, the chemical potential of the radicals is released explosively, since the reactions are exothermic and cause further heating. The result of such explosive chemistry (which has

been reproduced in the laboratory) is a sort of carbon-rich, refractory substance that is sometimes dubbed "yellow stuff" for lack of detailed knowledge about the chemical composition. Hence, after each dive into a molecular cloud a grain will end up without any ice but with its core surrounded by a mantle of *refractory organics*. This cycle of evolution was mainly explored by the American astrophysicist J. Mayo Greenberg.

The story ends in a particular molecular cloud — the one in which the Sun was formed. A certain part of it collapsed to form the solar system, and we may call this the *presolar cloud*. As shown in Fig. 2.17, the grains residing in the presolar cloud would have a structure involving both a silicate core, an organic refractory mantle, and an outer, icy mantle. Thus, each grain might have the composition of a typical comet, and comets may be aggregates of such grains (Greenberg and Hage 1990).

2.6.5. *Cometary material: The solar nebula model*

In the other scenario we start from the accretion disk around the nascent Sun, which is usually called the *solar nebula*. This formed

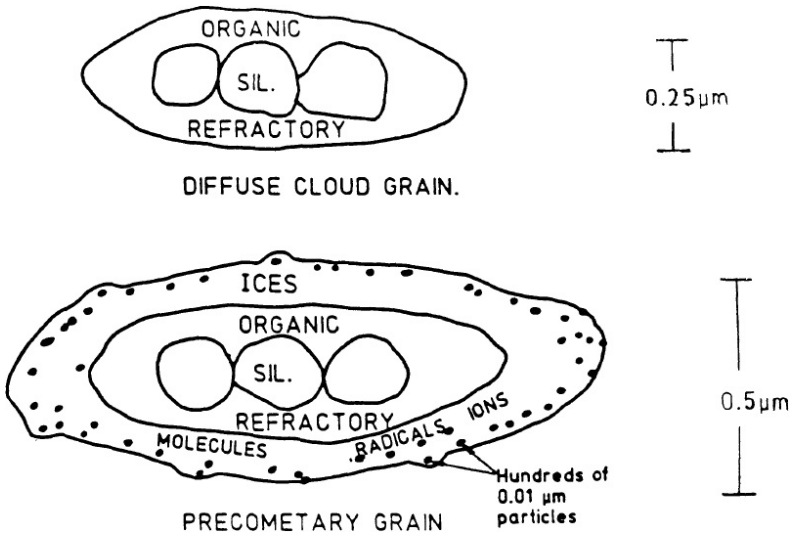


Fig. 2.17. Schematic illustrations of typical interstellar grains in two environments. The top picture shows a grain situated in a usual atomic hydrogen cloud ("diffuse cloud"), which has undergone one or more passages through molecular clouds and acquired a refractory organic mantle. The bottom picture shows such a grain during a subsequent stay in a molecular cloud, where an extra, icy mantle has been formed. Reproduced from Greenberg and Hage (1990) © AAS. Reproduced with permission.

out of the presolar cloud and should at least originally have contained the same type of grains as we described above. However, it is clear that the formation of the solar nebula involved substantial heating. For instance, meteorites often contain inclusions made of extremely refractory minerals, showing that there was a place and a time, where the temperature was high enough to keep only such minerals in the solid state. In the case of the CAIs, the condensation temperature is in excess of 1500 K.

The question is, what temperatures were reached in the part of the solar nebula, where comets grew? Only one thing seems certain, namely, that the organics were not vaporized, since it would be very unlikely for the complex, hydrogen-poor molecules to arise again after having been UV-dissociated in the gas phase. But there is no telling what happened to the ice, so it is worth considering the consequences, if the ice vanished by sublimation from the presolar grains and had to recondense in the solar nebula.

One constraint is given by the high abundance of oxidized species like CO and CO_2 in comets. When these molecules were set free to move in the solar nebula, which was dominated by H_2, they would fall victim to reduction into CH_4 and H_2O unless the density was low enough for these reactions to be kinetically inhibited (i.e., to have too long time scales). The fact that the gas-phase carbon in comets is oxidized rather than reduced thus sets a lower limit to the heliocentric distance of the comet-forming region.

If comets contain less water than the presolar grains did, as suggested by the Rosetta results, this should have some implications for our understanding of the way comets formed. H_2O molecules were available in such quantities that, when recondensation occurred, water ice could easily dominate over the refractory components. There would hence have to be some factor inhibiting this recondensation, at least partially. One possibility is that the time scale for grain growth and sedimentation was not very long compared to the time scale for the growth of icy mantles by condensation. This would more likely occur at smaller heliocentric distances, but the turn-over distance remains to be estimated by detailed calculations.

However, there is a different mechanism that seems established by observations and which may lead to a similar result. This is the above-mentioned radial mixing in the solar nebula. The high abundance of silicate minerals with high formation temperatures in the dust of comet 81P/Wild 2 can thus be explained, but if those minerals account for a large fraction of the entire silicate component, it also represents an important extra supply of refractories into the comet material.

2.6.6. *Cometary ice: Crystalline or amorphous?*

There are lots of solid phases of H_2O, but most of these are stable only at high pressures and are thus of no concern for comets. Here we deal with low temperatures and very low pressures, and then the options are few. Usual ice (I_h) is crystalline with a hexagonal crystal structure, and at low temperatures it may transform into cubic ice (I_c), which has a different crystal structure as the name implies. However, there is also an amorphous form of ice, which can be made in the laboratory by vapor deposition onto a cold plate at temperatures $T \sim 20$–$40\,\mathrm{K}$ or less.

The ice in comets is certainly not pure H_2O ice. Other molecules must also be present in the ice phase, since they too contribute to the outgassing. Different possibilities exist. The first to be proposed was *clathrate hydrates*. The clathrates are a class of compounds, in which "guest" atoms or molecules are imprisoned into the cavities offered by the lattice geometry of the crystalline ice. In hexagonal ice, one example often mentioned in the literature is methane clathrate ($CH_4{\cdot}^7H_2O$), since equilibrium models of solid formation in the solar nebula — when extended to very low temperatures — show that methane molecules would creep into the ice lattice at $T \sim 100\,\mathrm{K}$.

In the actual low-pressure conditions of the outer parts of the solar nebula, this is unlikely to proceed, but clathrates have nonetheless been popular. The reason is that the guest molecules are freed, as the ice sublimates, and one can thus explain why the usually observed coma species like C_2 and CN appear in comets approaching the Sun roughly at the distance of about 2.5 AU, where the H_2O ice starts to sublimate. If the parent molecules in question

would form their own ice, these would start sublimating at different distances. Note, however, that modern results show such coma species to originate largely from a semi-volatile component of comet material rather than volatiles frozen into ice.

Despite the popularity that clathrates once enjoyed, these are no longer a prime subject, when the chemistry of comets is discussed. It has been recognized since a long time that some comets outgas very large quantities of CO — much more than a clathrate could accommodate. Moreover, the fact that the outgassing patterns in comet Hale-Bopp shown in Fig. 2.5 show the minor species in general to follow the production curve of CO rather than that of H_2O is clearly contrary to the clathrate prediction.

In the framework of the interstellar model it is more tempting to think of the ice as amorphous, because the icy mantles on the cometary grains would have been deposited at temperatures similar to those of the laboratory experiments. Another attractive feature is the fact that amorphous ice has been found to be extremely rich in micropores, which form good trapping sites for guest molecules even in huge quantities (Bar-Nun and Kleinfeld 1989). From the work of Akiva Bar-Nun and his colleagues using a mixture of H_2O and other molecules for deposition, as illustrated in Fig. 2.18, we learn that the trapping efficiencies are very sensitive functions of the plate temperature (exponential fall-off), and below 40 K the amount of trapped gas even exceeds the amount of H_2O ice.

Amorphous ice is not absolutely stable. Upon slow heating, some release of trapped gas occurs already at very low temperatures due to annealing, but the main effect is due to crystallization, which proceeds at an exponentially increasing rate (Schmitt *et al.* 1989). Thus, at temperatures well below 100 K the crystallization time is millions of years or more, while at 130–140 K it is just a few hours. In addition, the settling of the H_2O molecules into a crystalline lattice is exothermic, so the ice gets heated and crystallizes even faster. This has led to the idea of episodic, explosive crystallization in comets occurring when the local ice temperature reaches the above-mentioned range, as described by Dina Prialnik and colleagues (e.g., Prialnik and Bar-Nun 1987).

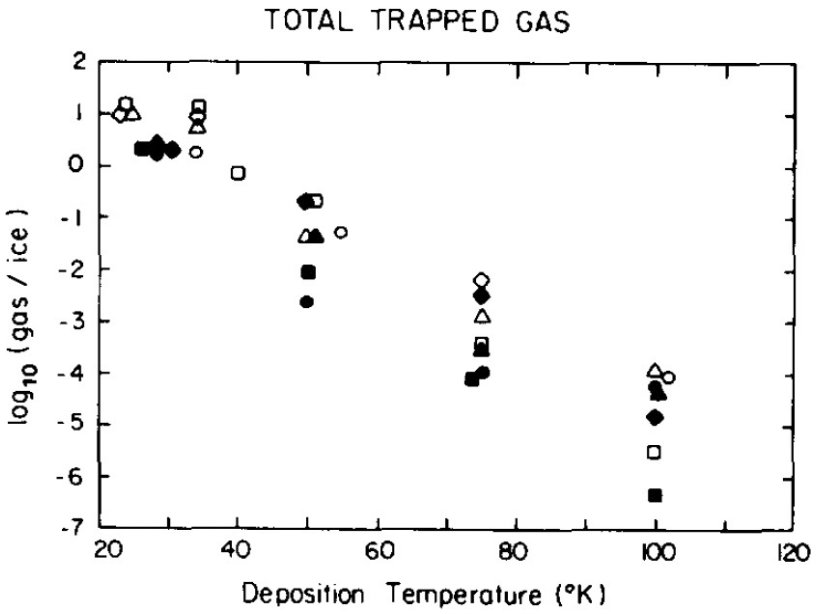


Fig. 2.18. Trapping efficiency of CH_4, CO, N_2 and Ar into amorphous ice, plotted with different symbols, versus the deposition temperature. Filled and open symbols denote different compositions of the flowed vapor. The plotted quantity is the log of the ratio between the amounts of trapped gas and ice. Reprinted from Bar-Nun, A. and Kleinfeld, I., *Icarus* **80**, 243–253 (1989), with permission from Elsevier.

Upon crystallization, most of the once trapped gases escape, as shown in Fig. 2.19. They are then free to diffuse through the pore system in either direction: toward the interior where they may recondense at lower temperatures, or toward the surface where they may flow into space — possibly dragging grains along. Thus, explosive crystallization offers a means to explain outbursts of activity sometimes seen in comets (see Sec. 4.4.3). It may also cause an increased activity of new comets from the Oort Cloud on their inward orbital branch (see Sec. 4.5.1). However, even though the amorphous-crystalline phase transition is certainly exothermic, evidently the released energy is not fully available for heating of the ice. It is possible that, depending on the amount of trapped gases, most or all of this energy may be consumed by unbinding the guest molecules from their trapping sites. In addition, most of the available energy will go into heating the dominant, refractory component of the comet material.

Still, there are reasons to be skeptic about cometary ice being amorphous. These have to do with the origin of the ice and the

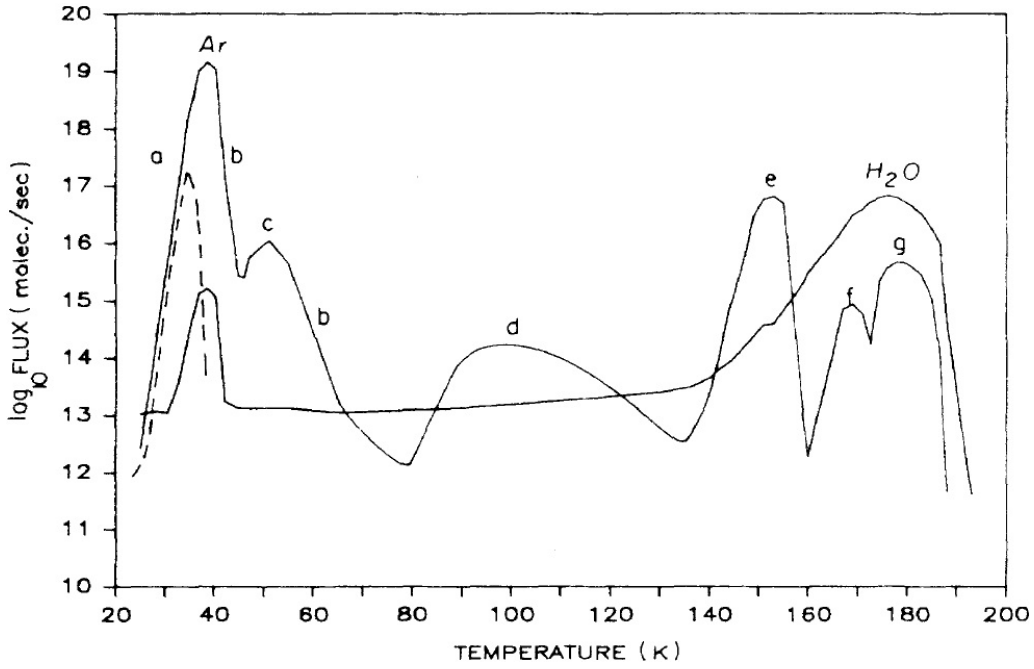


Fig. 2.19. Fluxes of argon and water from amorphous ice with trapped Ar, deposited at 20 K, versus the temperature during slow heating. Near 30 K frozen Ar sublimes from the ice surface, dragging along some H_2O. The next major peak in the Ar flux starts near 140 K and is due to crystallization of the ice. Near 180 K the ice sublimes. Reprinted figure with permission from A. Bar-Nun *et al.*, *APS Phys. Rev. B* **35**, 2427–2435 (1987). © (1987) by the American Physical Society.

thermal history of the cometary grains in the solar nebula. Although the formation of icy mantles on the primordial grains in the presolar cloud occurred at the same temperature as the cold plate, where amorphous ice was formed in the laboratory, there is an enormous difference in the ambient gas density, and thus the time scales of deposition are very different. Thus, while in the laboratory the deposited molecules have too little time to find their place in a crystal lattice, in the presolar cloud they might not have encountered this problem.

Second, if the presolar ice was indeed amorphous, it could possibly have crystallized due to shock heating, as the grains and their surrounding gas fell into the solar nebula at important velocities. It has even been proposed that the ice could have been vaporized, so that the extant cometary ice must have formed by condensation in the solar nebula under the ambient conditions. According to

calculations by Kouchi *et al.* (1994), this would rather result in crystalline ice throughout the region of comet formation.

In the presence of such uncertainties, one has to ask if there is any direct evidence of amorphous ice being responsible for cometary outgassing. It appears that such evidence may exist, but it is not unquestionable. One case in point is comet 29P/Schwassmann-Wachmann 1. With a perihelion distance of 5.7 AU and a discovery made in 1927, this comet is arguably the first discovered Chiron type comet. It has an orbit of very low eccentricity but is extremely variable in its activity. Most of the time it stays quiescent, but it undergoes sporadic, major outbursts with a very large increase in brightness (see Sec. 4.4.3).

Explosive ice crystallization has been proposed as the mechanism behind the outbursts of comet 29P, but there is no detailed, convincing argument in favor, so the issue is open. However, the low-level activity in the quiescent phase has been studied by radio astronomical observations with interesting results. These have revealed a persistent, anisotropic outflow of CO molecules from the nucleus plus a symmetric production of CO from grains surrounding the nucleus in a wide cloud (Gunnarsson *et al.* 2002). The nuclear source, accounting for about 1/4 of the molecules, is of particular interest. As seen in Fig. 2.20, this is a jet-like source emanating from the nucleus region facing the observer. Since the comet is always near opposition, this means a nuclear source near the subsolar point at all times.

It is obvious that there cannot be frozen CO on the nuclear surface, since this would disappear at once by sublimation. The source of the CO is therefore to be found below the surface. If it were located deeper than several rotational skin depths from the surface, the local temperature would stay almost constant, and the production of CO would occur on both day and night sides of the nucleus, which is not observed. Due to the low thermal inertia of comet material (Sec. 1.5), this skin depth is so small that the CO source must be very close to the surface. It is therefore natural to seek the explanation for the CO release among mechanisms that are triggered at the temperature of the 29P subsurface layer. This can be

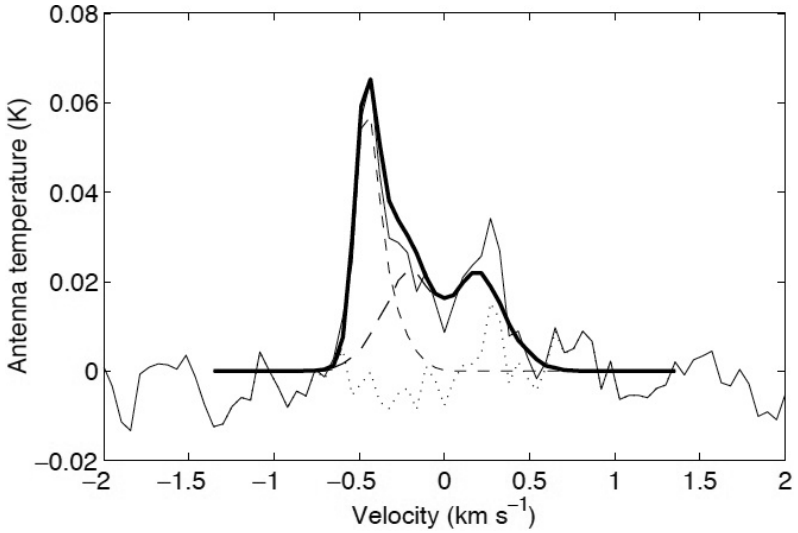


Fig. 2.20. Spectrum of comet 29P, observed in 1998 using the ESO/SEST telescope (thin, solid line). The emission is due to the 230 GHz CO ($J = 2 - 1$) transition. The thick line shows a best model fit, using a nuclear outgassing source (short dashes) and a distributed coma source (long dashes). The dots show the residual between model and observations. Reprinted from Gunnarsson, M. *et al.*, *Icarus* **157**, 309–322 (2002), with permission from Elsevier.

estimated at ∼130 K in good agreement with the H_2O crystallization temperature.

Hence, at any given time, the subsolar latitude of the nucleus would have a thin surface layer containing crystalline ice. Below this there is a crystallization front producing the CO outflow, which recedes at the same average rate as the surface is eroded by the release of icy grains. The crystallization front must be prevented from explosive recession, and thus very little of the latent heat of crystallization can be available for heating the ice. The grains seen in the 29P coma would consist of this material and might produce H_2O (though still unseen) by sublimation due to the low albedo of the grains. This seems like a good working model to study the matter further, but there is no proof that it represents the truth.

Chapter 3

Comet Dynamics

As we shall see, comet dynamics is a rather heterogeneous subject, and different features are important for different groups of comets. There is no common denominator, but in nearly all the situations, comets move on heliocentric orbits that are relatively stable for significant amounts of time. The two-body problem (Sun–comet) is then a zero-order approximation, to which one may add small perturbations — typically, due to the gravity of the planets.

These perturbations induce changes of the heliocentric orbital elements. Usually, when the comet is not very close to a planet, the changes are slow. However, close approaches may often occur, and these are typically of short duration but may change the heliocentric orbit dramatically. In its general formulation, the N-body problem of the Sun, several planets and the comet is not fully integrable, and this holds even for the 3-body problem, where only one planet is involved. However, under certain conditions, the latter problem allows us to derive an analytic formulation of the comet's energy of motion, which is independent of time and thus serves to constrain the way the orbit may evolve.

3.1. Circular Restricted Three-body Problem

We consider the theoretical problem of the motion of three bodies, where one of these is so small that it has no influence at all on the motions of the other two. This is the *restricted 3-body problem*, and it deals with the motion of the small body, because the other two

move on unperturbed orbits around their center of mass. These orbits are typically assumed to be either elliptic or circular. In the elliptic problem, there is no closed expression constraining the small body's motion. However, the *circular, restricted 3-body problem* allows the formulation of an energy integral. This is of great use in comet dynamics, where as a first approximation one considers the comet as massless and the Sun and Jupiter to represent the rest of the Universe with Jupiter's orbit around the Sun being circular. In reality, this is quite a good approximation, so the energy integral serves as a good hint on the orbital evolution of the comet.

Briefly, the derivation of the energy integral goes as follows. We place ourselves in a rotating coordinate system, where the two massive bodies are at rest. The acceleration of the massless body is now due to the gravities of the massive bodies and the centrifugal force. All these components are conservative, and the massless body moves in a fixed potential. Choosing the units of mass, distance and time such that the total mass of the system is unity, the distance between the massive bodies is unity, and the angular velocity of the circular motion is also unity, we can at once write:

$$C = -2E = (x^2 + y^2) + 2\left\{\frac{1-m}{\rho_1} + \frac{m}{\rho_2}\right\} - v^2, \qquad (3.1)$$

where we have taken a rotating (x, y, z) frame with the z-axis parallel to the angular momentum of the 2-body system, m is the smaller of the two masses, ρ_1 and ρ_2 are the distances of the massless body from the other two bodies, and v is the velocity of the massless body in the rotating system. Note that with the units chosen, the gravitational constant also has the value unity. The quantity E is the constant energy of motion per unit mass of the small body, and C goes under the name *Jacobi integral*, having been first derived by C. G. J. Jacobi in 1836.

Allowing for the factor -2, the term x^2+y^2 in Eq. (3.1) represents the centrifugal energy, the terms in the curly brackets represent the potential energy, and the last term represents the kinetic energy. In 1889, F.F. Tisserand used the Jacobi integral to derive a criterion to

look for the possible identity of comet pairs, where one was observed before and the other after an encounter with Jupiter, which led to a large change of the orbital elements (the so-called *Tisserand criterion*). This is a simple expression in terms of the orbital elements, which would be quasi-constant in the circular problem as long as the comet moves far from Jupiter, and it reads

$$T = \frac{a_J}{a} + 2\sqrt{\frac{a}{a_J}(1-e^2)}\cos i. \tag{3.2}$$

Here a_J is the semi-major axis of Jupiter's orbit, and a, e and i are the semi-major axis, eccentricity and inclination of the comet orbit. The latter should in principle be measured with respect to Jupiter's orbital plane, but in practice the ecliptic inclination can be used as well. With the units chosen for the Jacobi integral, we of course have $a_J = 1$.

The quantity T is called the *Tisserand parameter*, and it means the same as C. It is rarely used for the purpose it was conceived for, but as seen in Sec. 1.4, it plays an important role when classifying comets in terms of their orbits. It is only approximately constant, because the ellipticity of Jupiter's orbit may cause changes — in particular during close encounters — and the perturbations due to the other planets have no reason to leave T unaffected. In principle, each planet can in theory be equipped with its own Tisserand parameter, which would be useful for comets that are under that planet's control. To allow for this possibility, one sometimes writes T_J instead of T to avoid ambiguity. This was done in Sec. 1.4, when the Tisserand parameter was used for orbital classification of comets.

However, in practice T is used mainly for short-period comets, and these are typically under Jupiter's control and rarely affected significantly by any other planet. Together with the relative smallness of Jupiter's eccentricity (0.048), this means that the orbital evolutions experienced by short-period comets are in fairly good agreement with the predictions of Eq. (3.2) using constant T. In Table 3.1 we list several large orbital transformations undergone by Jupiter Family comets in connection with close encounters with Jupiter. It is seen

Table 3.1. Examples of major orbital transformations experienced by numbered periodic comets. Abbreviations used are: 'Ch-G' for Churyumov-Gerasimenko, 'S-Ch' for Smirnova-Chernykh, and 'W-K-I' for West-Kohoutek-Ikemura. The year of closest approach to Jupiter and the minimum jovicentric distance ($\Delta_{\min}$) attained are listed. For comets 74P and 82P the encounters were long-lasting and involved two perijove passages. The perihelion (q) and aphelion (Q) distances and the Tisserand parameter (T) are given before and after the respective encounters, separated by a hyphen.

Comet	Year	$D_{\min}$ (AU)	q (AU)	Q (AU)	T
16P/Brooks 2	1886	0.001	5.47–1.95	14.48–5.43	2.98–2.89
39P/Oterma	1937	0.165	5.79–3.39	8.06–4.56	3.02–3.03
39P/Oterma	1963	0.095	3.39–5.45	4.56–8.99	3.03–3.00
67P/Ch-G	1959	0.052	2.76–1.29	5.90–5.73	2.76–2.75
74P/S-Ch	1955–1963	0.245–0.467	5.65–3.55	11.92–4.80	3.01–3.01
76P/W-K-I	1972	0.012	4.93–1.40	13.45–5.29	2.73–2.68
81P/Wild 2	1974	0.006	4.97–1.48	21.53–5.23	2.77–2.88
82P/Gehrels 3	1970–1973	0.001–0.041	5.68–3.43	8.19–4.67	3.02–3.03

that T was not much perturbed in spite of dramatic changes of the standard orbital elements. The changes that did occur were typical of those that Jupiter's orbital eccentricity induces during very close or long-lasting encounters.

The comets of the Jupiter Family are characterized by low inclinations (see Fig. 1.4), and they generally have $\cos i \approx 1$. A co-planar orbit ($i = 0$) will remain co-planar when perturbed by Jupiter, and in general, low-inclination orbits also stay with low inclinations. It is therefore of interest to use Eq. (3.2) to draw evolutionary curves in the (a, e) plane for constant T with $i \equiv 0$. In Fig. 3.1 we see a rendering of such a diagram using the aphelion and perihelion distances (Q and q) on the axes.

We see that the observed Jupiter Family comets are likely to have evolved into their current orbits from a source population with larger perihelion distances and larger aphelion distances as well. As the development in discovery techniques has allowed comets with larger perihelion distances to be discovered, our records thus start to probe deeper into this source population. However, the diagram concentrates on the vicinity of the observed Jupiter Family and does

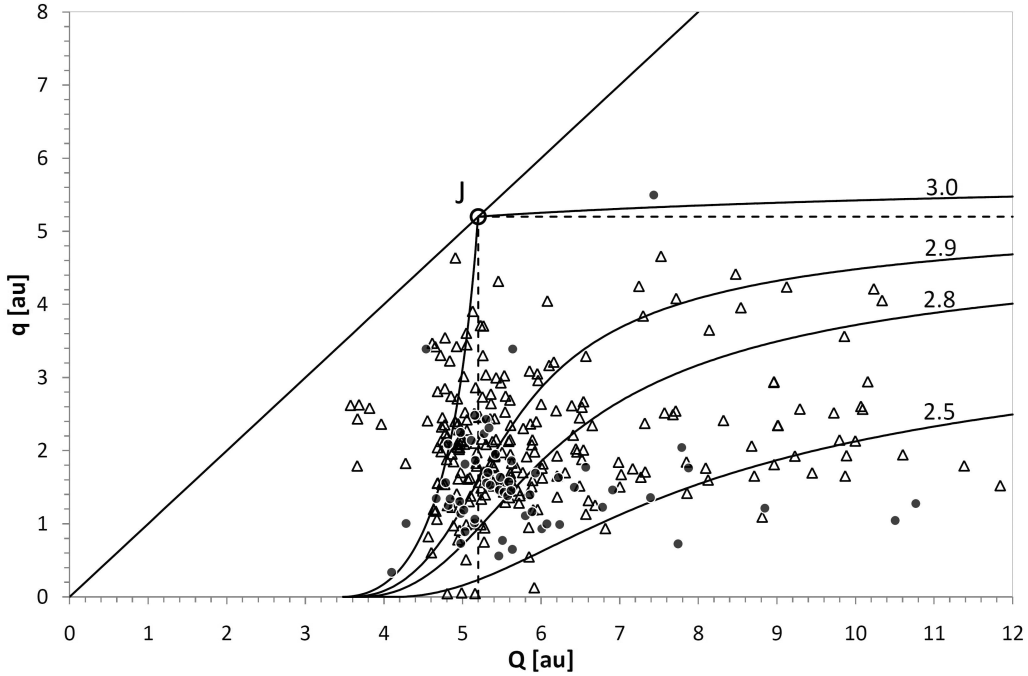


Fig. 3.1. Level curves of constant T for zero inclination, plotted in a diagram of perihelion versus aphelion distance. The symbols refer to the orbits of the numbered periodic comets. Filled circles denote comets discovered before 1950, and open triangles denote more recent discoveries. Courtesy T. Wiśniowski.

not allow to gain insight into the ultimate source of the comets. This issue will be dealt with in Sec. 5.5.

There are some interesting differences between the evolutionary curves in Fig. 3.1. In the circular 3-body problem, only comets with $q < a_J$ or $Q > a_J$ can have very close encounters with Jupiter. But for $T \approx 3$ the curves lead into orbits that do not satisfy this requirement, so the question arises if such evolutions are possible. In practice, they are possible due to the ellipticity of Jupiter's orbit together with the fact that slow encounters may bring the comets close to Jupiter even if the comet motions upon approach aim relatively far from the planet (see Sec. 3.2.2).

In fact, as shown by Ernst Öpik (1951), the value of T is directly related to the speed of approach. Suppose the comet to be situated close to Jupiter's orbit. If we consider Eq. (3.2) with $a_J = 1$, according to the vis-viva integral, the term $1/a$ equals $2 - V^2$, where V is the speed of the comet in a fixed frame. The second term can be identified with twice the component of the comet's angular momentum perpendicular to Jupiter's orbital plane, which can also

be written $L_z = V_{\rm tr}$, where $V_{\rm tr}$ is the comet's velocity component parallel to Jupiter's velocity in the fixed frame upon encounter. In the normalized units, the Tisserand parameter can therefore be expressed as

$$T = 2 - V^2 + 2V_{\rm tr}. \tag{3.3}$$

The relative encounter velocity U between the comet and Jupiter can be written

$$U^2 = (\mathbf{V} - \mathbf{C})^2 = V^2 + 1 - 2\mathbf{V} \cdot \mathbf{C} = V^2 + 1 - 2V_{\rm tr}, \tag{3.4}$$

where $\mathbf{C}$ is Jupiter's velocity vector ($C = 1$).

From Eqs. (3.3) and (3.4), we obtain

$$U^2 = 3 - T. \tag{3.5}$$

Here we see that the speed of approach (U) can have a real value only if $T \leq 3$, and for $T = 3$, the speed vanishes. This is consistent with what was said above, namely, that close encounters are formally impossible in the circular problem for orbits with $T > 3$. We also see that comets that encounter Jupiter with $T \approx 3$ will do so with a very small velocity, so that Jupiter's gravity may attract them even from large distances.

Finally, let us note that the definition of Jupiter Family comets in Sec. 1.4 essentially means that these comets would approach Jupiter with speeds less than Jupiter's own orbital velocity, while the opposite is the case for Halley-type comets.

3.2. Close Encounters

Table 3.1 showed several cases, where the orbits of Jupiter Family comets were changed more or less dramatically as a result of close encounters with Jupiter. But the recent dynamical history of the entire Jupiter Family is full of such encounters and their resulting jumps in orbital element space. Only a minority of comets have been spared of these events. Thus, for the Jupiter Family it is clear that close encounters play a dominant role in the orbital

evolution. Moreover, the dynamics of long-period and Halley-type comets features orbital jumps at close encounters as an important phenomenon as well. Let us therefore take a closer look at the workings of the close encounters.

The fact that the heliocentric orbit may change dramatically means that the approximation of elliptic motion breaks down during the close encounter. Hence, the planet is surrounded by a *sphere of influence*, inside which it replaces the Sun as the main arbiter of the comet's motion. A rough estimate of the size of the sphere of influence can be obtained by comparing the magnitudes of the forces exerted on the comet by the Sun and the planet. In the heliocentric frame the central force is approximately $GM_\odot/a_p^2$, if the comet is situated in the vicinity of the planet (a_p is the semi-major axis of the planetary orbit). The perturbing force by the planet is GM_p/Δ^2, where Δ is the distance between the comet and the planet. Equality of the two forces occurs for

$$\Delta = R_\mathrm{h} = a_p m_p^{1/2}, \tag{3.6}$$

where $m_p = M_p/M_\odot$. As long as $\Delta > R_h$, the central force is larger than the perturbing force, and the heliocentric motion is relatively stable, but when the opposite inequality holds, the heliocentric ellipse cannot be expected to describe the comet's motion even approximately.

If we instead place ourselves in the planetocentric frame, we can ask the same question for the two-body motion of the comet under the gravity of the planet. In this case, the central force is GM_p/Δ^2, while the perturbing force due to the Sun is a tidal force, which can be expressed as $GM_\odot\Delta/a_p^3$. Equality of the two forces now occurs for

$$\Delta = R_\mathrm{p} = a_p m_p^{1/3}. \tag{3.7}$$

Hence, the planetocentric motion can be approximately described as a conic section as long as $\Delta < R_\mathrm{p}$ but not for larger Δ values.

Both these spheres around the planet, with radii R_h and R_p, can be called spheres of influence. Obviously, $R_\mathrm{h} < R_\mathrm{p}$, so there is a range of Δ between the two radii, where both the heliocentric and

planetocentric two-body orbits are relatively useful to describe the motion. If we take Jupiter as the planet, we have $R_\mathrm{h} \simeq 0.16\,\mathrm{AU}$ and $R_\mathrm{p} \simeq 0.52\,\mathrm{AU}$.

3.2.1. *Hyperbolic deflection*

Let us first assume that the comet approaches the planet with a non-negligible speed. To be precise, we take this speed to be large enough for the planetocentric orbit to be hyperbolic. This is the usual situation, and the peculiar case where it does not apply will be treated afterwards. Of course, an accurate computation of the outcome of a close encounter cannot be based on the two-body approximation, but the model of a close encounter as a *hyperbolic deflection* is still useful.

In practice, the hyperbolic deflection can be realized as follows. The comet moves along an osculating elliptic orbit around the Sun at the moment, when it enters into the planet's sphere of influence. Its planetocentric velocity vector at that time can be written $\mathbf{U} = UV_p\hat{\mathbf{U}}_\mathrm{in}$, where U was defined in Sec. 3.1, V_p is the orbital speed of the planet in arbitrary units, and $\hat{\mathbf{U}}_\mathrm{in}$ is the unit vector defining the direction of $\mathbf{U}$. At the given entry point at planetocentric distance R_p, $\mathbf{U}$ defines a hyperbolic orbit around the planet. The comet is assumed to follow this orbit until it reaches the surface of the sphere of influence again on the outward branch of the hyperbola. At this moment, the planetocentric velocity vector $\mathbf{U}' = U'V_p\hat{\mathbf{U}}_\mathrm{out}$ is used to compute a new, osculating heliocentric orbit. It is not strictly necessary to use R_p for the radius of the sphere of influence — any value down to R_h is equally admissible.

From the properties of hyperbolic motion, we realize that $U' = U$. Now, consider the approximation that this motion takes no time at all. This means that the perturbation of the heliocentric orbit is an instantaneous change of the velocity vector $\mathbf{V}$, which is triggered by an instantaneous rotation of the planetocentric direction of motion $\hat{\mathbf{U}}$ — i.e., the hyperbolic deflection. We note that this perturbation will leave T unaffected in accordance with Eq. (3.5).

To determine the amount of deflection, let us consider the so-called *b-plane* (Valsecchi *et al.* 2003), which contains the center

of the planet and is perpendicular to $\hat{\mathbf{U}}_{\rm in}$. We approximate the actual deflection, which occurs over a finite part of the hyperbola, by the full deflection between the two asymptotes, so that $\hat{\mathbf{U}}_{\rm in}$ defines the direction of the incoming asymptote. We also take UV_p to be the hyperbolic velocity at infinity, the so-called unperturbed encounter velocity. The point in the b-plane, to which the comet aims, is situated at distance b from the center of the planet, and for this we use the term *impact parameter*.

The plane of the hyperbola is perpendicular to the b-plane. Denoting the angle between its asymptotes by $\pi - \gamma$, we have

$$\tan\frac{\gamma}{2} = \frac{GM_p}{b(UV_p)^2}. \tag{3.8}$$

This shows that the deflection angle (if small) is directly proportional to the mass of the planet and inversely proportional to the impact parameter and the square of the unperturbed encounter velocity.

The instantaneous change of $\mathbf{V}$ implies changes in the orbital energy and angular momentum of the comet. To describe these changes, it is useful to consider the angle θ between the vectors $\mathbf{U}$ and $\mathbf{V}_p$, as illustrated in Fig. 3.2. When describing the orientation of $\mathbf{U}$ with respect to $\mathbf{V}_p$, θ is the polar angle, and there is also an azimuthal angle ψ, which is not shown in the figure. The angle between the in- and outgoing vectors $\mathbf{U}$ and $\mathbf{U}'$ is γ, and both vectors

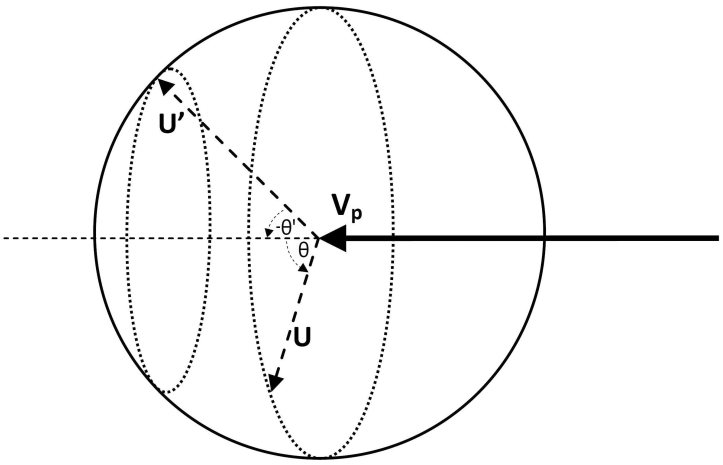


Fig. 3.2. The geometry of velocity space at a close encounter, using vectors that are defined in the main text. The dotted ellipses represent latitude circles at different co-latitudes θ and θ'. Courtesy T. Wiśniowski.

are situated in the *scattering plane*. However, in general, $\mathbf{V}_p$ is not situated in this plane. Hence, $\psi' \neq \psi$, and $\gamma \neq \theta - \theta'$.

Referring the comet's velocity vector approximately to the position of the planet, the orbital energy is given by the vis-viva integral as

$$E = -\frac{GM_\odot}{2a} = \frac{1}{2}V^2 - \frac{GM_\odot}{r_p}, \tag{3.9}$$

where r_p is the orbital radius of the planet. Using the cosine theorem, we can write

$$E = \frac{1}{2}V_p^2(1 + U^2 + 2U\cos\theta) - \frac{GM_\odot}{r_p}, \tag{3.10}$$

before the encounter and the same expression with θ replaced by θ' after the encounter. We thus have:

$$\Delta E = UV_p^2(\cos\theta' - \cos\theta). \tag{3.11}$$

For the perturbation of the angular momentum, we have

$$\Delta L = r_p \cdot UV_p(\cos\theta' - \cos\theta). \tag{3.12}$$

Using Eqs. (3.11) and (3.12), the perturbations of inverse semi-major axis and eccentricity can easily be calculated from the values of U, θ and θ'. It is easy to realize that the expression $\cos\theta' - \cos\theta$ takes its largest value for a given γ, if $\mathbf{V}_p$ is situated in the scattering plane and $\theta' = \theta - \gamma$. In this case, for the inverse semi-major axis we can derive:

$$\Delta\left(\frac{r_p}{a}\right) = 2U\{\cos\theta - \cos\theta'\} = 4U\sin\frac{\gamma}{2}\sin\left(\theta + \frac{\gamma}{2}\right). \tag{3.13}$$

Hence we see that, for a given impact parameter, encounters with very low velocity ($U << 1$) yield very small energy perturbations. From this, we may expect that the speed of transfer along the evolutionary curves in Fig. 3.1 caused by close encounters with Jupiter is reduced for Tisserand parameters approaching the critical value of 3. However, this is only valid as long as the encounter speeds are large enough to warrant the approximation of instantaneous, hyperbolic deflection. Thus, let us also consider the other extreme

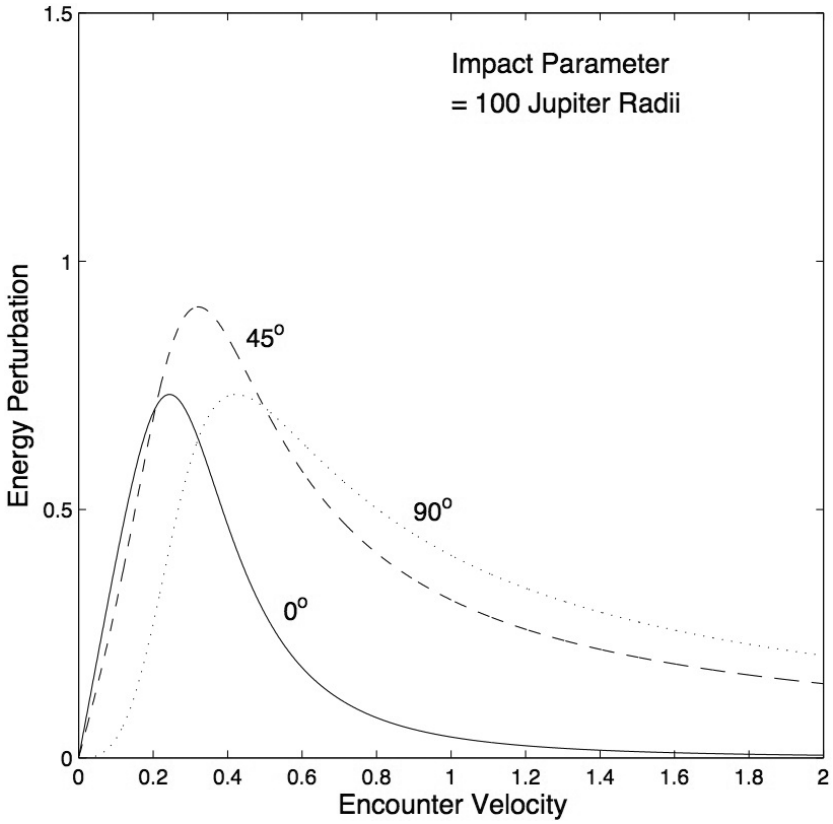


Fig. 3.3. Perturbations of a_J/a, computed with Eqs. (3.8) and (3.13), for a common impact parameter $b = 100R_J$ and three values of the angle θ characterizing the approach direction. Credit: H. Rickman, "Cometary Dynamics," in Lecture Notes in Physics, Vol. 790 (2010), pp. 341–399, © Springer-Verlag Berlin Heidelberg 2010. With permission of Springer Nature.

case of extremely large velocity ($U \to \infty$), for which Eq. (3.8) tells us that $\sin\frac{\gamma}{2} \to 0$ proportional to U^{-2}. Consequently, the energy perturbation again tends to zero.

This behavior is illustrated in Fig. 3.3, which uses three different approach directions with a common impact parameter of 100 jovian radii for encounters with Jupiter. The largest energy perturbations are generally obtained with encounter velocities $U \simeq 0.3$–0.4, corresponding to Tisserand parameters $T \simeq 2.8$–2.9. This is the range that is most frequently populated by observed Jupiter Family comets.

The method of calculating orbital perturbations induced by close encounters with a planet, which was just described, was introduced by Öpik (1951). It has been found to be very useful as a tool for statistical investigations into the dynamical evolutions of objects that are subject to such encounters, like comets and early solar

system planetesimals. In addition, it serves well to provide an insight into how the encounters work, which is not available when tracing the motions by numerical integrations. It is not necessary for the encounters to be very deep or very fast — the only scenario where the method fails miserably is the one of very slow encounters. Otherwise, it is not very accurate but at the same time not deceptive.

3.2.2. *Slow encounters*

Let us now face the problem of the very slow encounters. These can be discussed with the aid of a concept called *zero-velocity surfaces*, which relates to the Jacobi integral. In Eq. (3.1), let us put $v^2 = 0$, which means that the comet is at rest in the rotating frame. For any particular value of C, the equation then defines a three-dimensional surface in (x, y, z) space, called the zero-velocity surface. This can be approached by the comet, but only at a vanishing speed, and it can never be crossed.

If C surpasses the critical value of 3 by a significant amount, the comet may move in three domains: far outside the planetary orbit, within a large ovoid around the Sun, or inside a small ovoid enclosing the planet. These domains are disconnected. However, if we let C decrease toward 3, the zero-velocity surfaces change shape and new situations may occur. At $C = 3.0388$, the two ovoids meet at the inner *Lagrangian point* on the line between the planet and the Sun, usually denoted L_1. For Jupiter, this is situated 0.35 AU from the planet. For slightly smaller values of C, there are two allowed domains of motion: the exterior one, and a common one formed by the merger of the heliocentric and planetocentric ovoids. At $C = 3.0375$, this common domain meets the exterior one at the outer Lagrangian point L_2, which for Jupiter is situated 0.36 AU from the planet. For slightly smaller values of C, comets are free to move between the inner and outer heliocentric regions by passing at small velocity through the ovoid-like zone between the Lagrangian points.

In Fig. 3.4 we show a cut of two zero-velocity surfaces with Jupiter's orbital plane, yielding the dashed and full-drawn curves. The dashed, quasi-elliptic curve represents the largest jovicentric region, from which there is no escape in the circular restricted 3-body

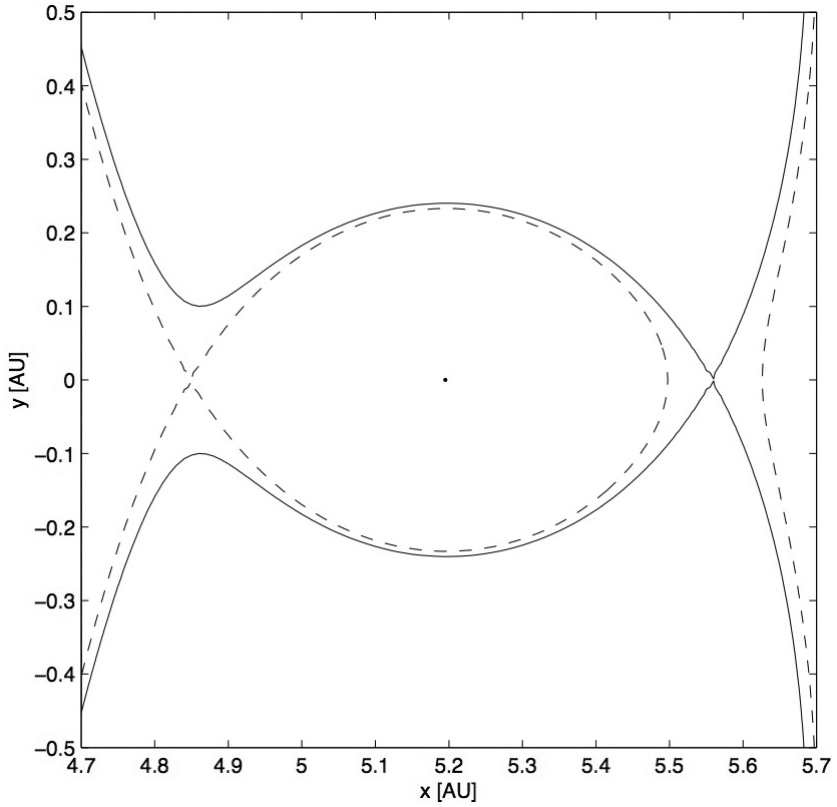


Fig. 3.4. Zero-velocity curves in Jupiter's orbital plane for two values of the Jacobi constant, passing through the inner (dashed) and outer (full-drawn) Lagrangian points. The Sun is at the origin, and Jupiter is at $x = 5.2$ AU. Credit: H. Rickman, "Cometary Dynamics," in Lecture Notes in Physics, Vol. 790 (2010), pp. 341–399,

problem. Such a region exists around any planet and is usually referred to as the *Hill sphere* due to its roughly spherical shape. Its radius is the distance from the planet to the L_1 point and can be shown to be

$$r_H = r_p \left(\frac{m_p}{3} \right)^{1/3}, \tag{3.14}$$

where m_p is the mass of the planet expressed in solar masses.

The surface of the Hill sphere can be seen as a stability limit for planetocentric satellite motion. Within the circular restricted 3-body problem, as mentioned, there is no escape for objects moving inside this limit, but there is also no access into the Hill sphere for objects moving outside. The fact that reality in the solar system differs from this idealized problem has important consequences in comet dynamics. Imagine a comet that approaches Jupiter from the solar

or anti-solar direction with a C value close to the critical range as just described. If C is such that there is a gap at the Lagrangian point in question, this gap would remain forever in the idealized problem, but in reality (due mostly to the eccentricity of Jupiter's orbit and the perturbations by Saturn) the gap may temporarily close — only to open up again at a later time. This allows for the possibility of *temporary satellite captures* around Jupiter, when a comet may spend an extended period of time orbiting within Jupiter's Hill sphere and may in fact be gravitationally bound to the planet. Such captures in the near past or future are known for several Jupiter Family comets, and an example from orbital integrations by Tancredi *et al.* (1990) is shown in Fig. 3.5.

Clearly, a comet may enter into Jupiter's Hill sphere from one side and exit on the other independent of the details of the intervening motion. In such a case, the heliocentric orbits before and afterwards are very different from each other. Due to the requirement of a

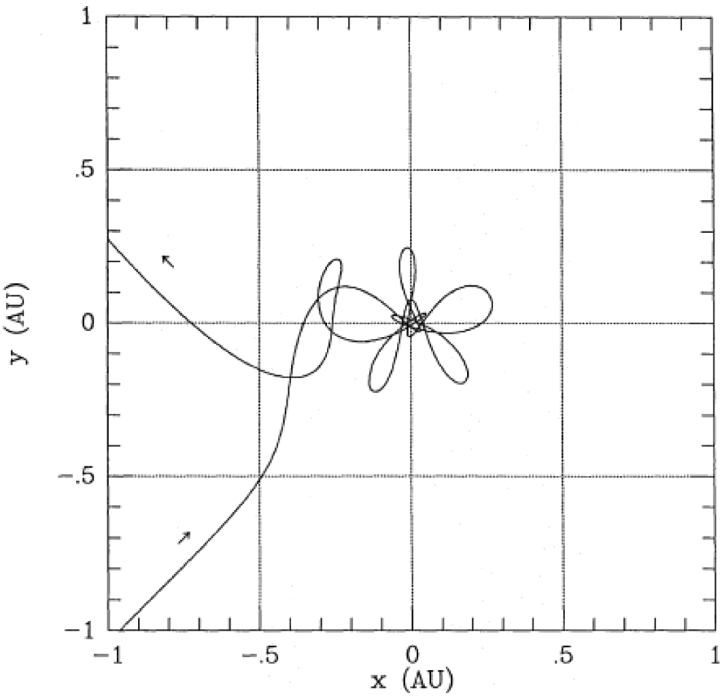


Fig. 3.5. Trajectory of comet 111P/Helin-Roman-Crockett with respect to Jupiter (plotted at the origin) during a temporary satellite capture predicted to occur around the year 2075. The frame rotates so that the Sun is always on the negative x-axis. Both entry and exit occur close to the L_1 point. Credit: G. Tancredi *et al.*, *A&A* **239**, 375–380 (1990), reproduced with permission © ESO.

very slow motion relative to Jupiter, the heliocentric radial velocity must be small, and the comet is either close to perihelion or close to aphelion. Thus, as explained by Carusi and Valsecchi (1979), there is a route between comet orbits that are tangent to Jupiter's at perihelion or aphelion via low-velocity encounters, possibly long-lasting and involving temporary satellite captures.

When comets exit from Jupiter's vicinity via the L_1 gap after low-velocity encounters, they enter into a special kind of heliocentric orbits, as explained by Tancredi *et al.* (1990). These orbits have aphelia somewhat inside the L_1 point and periods close to 2/3 that of Jupiter. Since this resonance is also shared by the Hilda group of asteroids, the comets in question are often called *quasi-Hildas*.[1] Contrary to the asteroids, these have no protection from close encounters with Jupiter, and numerical tracing of the motions of real quasi-Hildas reveals that they are just temporary visitors into the Jupiter Family.

Comet 39P/Oterma is a case in point. This comet had a long-lasting encounter with Jupiter with closest approach in 1937, leading from an outer orbit with $q \simeq 5.8\,\mathrm{AU}$ and $Q \simeq 8.1\,\mathrm{AU}$ into an inner orbit with $q \simeq 3.4\,\mathrm{AU}$ and $Q \simeq 4.6\,\mathrm{AU}$. It was discovered in 1943 by Finnish astronomer Liisi Oterma and kept under observation in the 1940s and 1950s while orbiting three times around the Sun. Then a new slow encounter with Jupiter with closest approach in 1963 transferred the comet back into an outer, Chiron-type orbit with $q \simeq 5.5\,\mathrm{AU}$ and $Q \simeq 9.0\,\mathrm{AU}$, and observations ceased until 2001, when the comet was rediscovered by a team led by American astronomer Yanga Fernández. The two orbital transformations were listed in Table 3.1.

3.3. Lidov–Kozai Cycles

While the observable comets are almost always susceptible to close encounters with Jupiter in the long run, it is worth considering the likelihood of a close encounter for a particular set of orbital elements.

[1]These are marked with sky-blue dots in Fig. 1.5.

Generally speaking, this depends on two factors (Rickman *et al.* 2014). First, one has to determine the smallest distance between the elliptical orbits of the comet and Jupiter to see if a close encounter is at all possible. This involves a calculation of the *MOID* (minimum orbit intersection distance; see Bowell and Muinonen 1994), for which a numerical method has been described and made public by Wiśniowski and Rickman (2013). Second, if the MOID is small enough to allow a close encounter, one needs to evaluate the probability of both objects passing their MOID points close enough in time for a close encounter to occur. Such a probability was analyzed for the case of an actual collision, assuming random elements, by Rickman *et al.* (2014), but the method is easily applicable to close encounters with Jupiter or any other planet.

The stochastic approach to the timing problem has obvious drawbacks for Jupiter Family comets because of the influence of mean motion resonances, and to some extent, the same is true for Halley Type comets too. We will now discuss *secular perturbations* in the absence of close encounters, focusing primarily on long-period comets. These are favored by their much higher average inclination, which tends to keep the MOID values large for most of the time.

Secular perturbations mean long-term variations of the heliocentric orbital elements under the influence of a perturbing planet. By long-term variations we mean that the orbital position of the comet does not matter, and the differential equations governing these variations are averaged over the orbital periods of the comet and the planet, which then acts like an elliptical arc of matter spread along its actual orbit. A first approximation is offered by a linear solution, where the semi-major axis of the comet remains constant. The two orienting angles ϖ (longitude of perihelion) and Ω (longitude of the ascending node) are then linear functions of time characterized by their secular frequencies. The eccentricity oscillates with the frequency of ϖ, and the sine of the inclination oscillates with the frequency of Ω.

These secular frequencies are the rates of circulation of the apsidal and nodal lines, respectively. *Secular resonances* occur, when either of these rates coincides with the rate of the perturbing planet.

The linear approximation then breaks down, and the eccentricity or inclination may undergo much larger variations. This phenomenon is of importance in the asteroid belt and, at least potentially, for Jupiter Family or Halley Type comets, but not for long-period comets. Halley Type and long-period comets may experience a different type of resonance, which can be visualized as follows.

Over a long enough period of time, the planet's apsidal and nodal lines circulate so that it acts approximately like a circular annulus of matter in the invariable plane,[2] centered on its constant semi-major axis. In this situation, neither ϖ nor Ω is important for the secular evolution of the comet orbit. In fact, the comet experiences the attraction of a circular annulus of matter in the fixed, invariable plane. This perturbing force cannot change the comet's orbital energy, and since it is confined by circular symmetry to the meridional plane spanned by the normal vector of the invariable plane and the comet's heliocentric radius vector, the associated torque does not change the comet's angular momentum component perpendicular to the invariable plane. Thus, the expression $L_z = \sqrt{a(1-e^2)}\cos i$ remains constant, as does a.

However, the perturbing force and its torque do perturb the comet orbit, and this happens by changing the total angular momentum vector $\mathbf{L}$. Due to the constancy of L_z and a, there are coupled variations of e and i, which are associated with the absolute value and direction of $\mathbf{L}$, respectively. These variations relate to the variation of the only angular element that matters, namely, the argument of perihelion $\omega = \varpi - \Omega$. Depending on ω, the nodes of the comet orbit may have different positions with respect to the planetary annulus, and this governs the variation of $\mathbf{L}$.

The ω variation may be a circulation, in which case ϖ and Ω circulate at different rates. Alternatively, ω may librate around $\pm\pi/2$, in which case there is a 1:1 resonance between the rates of ϖ and Ω. In the (L, ω) parametric plane, trajectories near the separatrix between these two modes exhibit large variations of L

[2]This is a plane perpendicular to the constant, total angular momentum of all the planetary orbits.

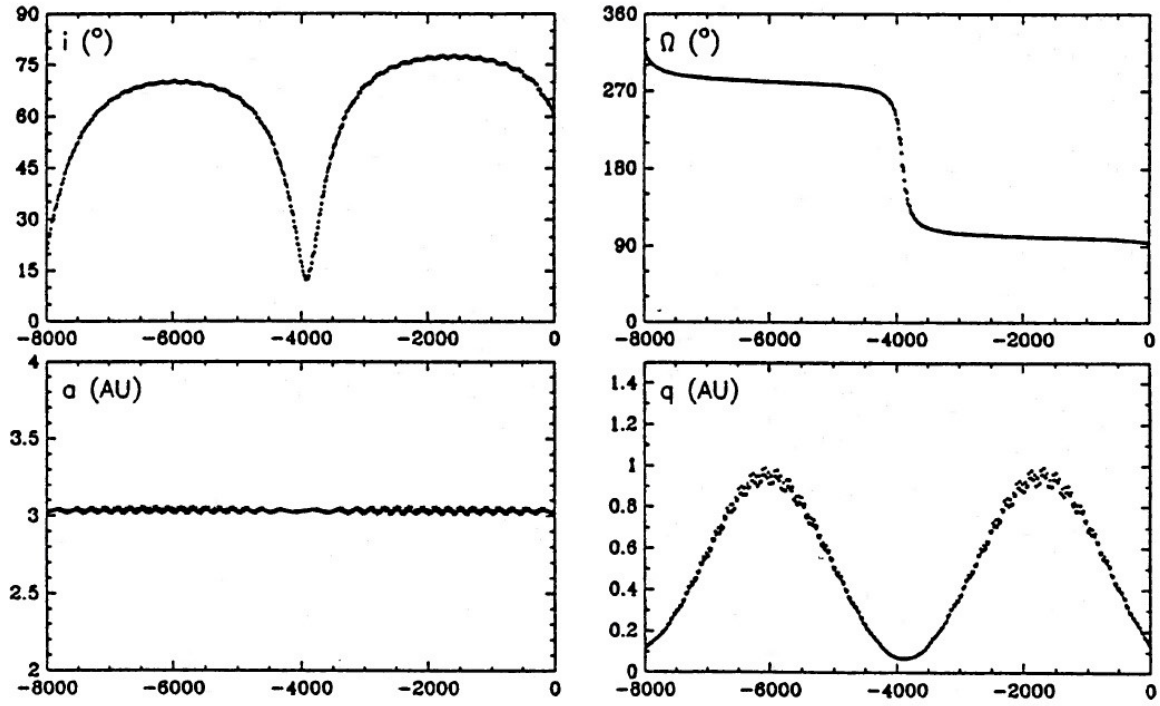


Fig. 3.6. Long term variations of the inclination (i), longitude of the ascending node (Ω), semi-major axis (a) and perihelion distance (q) of comet 96P/Machholz 1. The time scale is in years and goes backward from a zero-point in 1986. Reproduced from Gonczi *et al.* (1992) with permission.

and thereby of e and i, especially when L_z is small. As required by the constancy of L_z, these variations are in anti-phase, so that i has its minimum when e is at maximum and the converse. Alternatively, the variations of perihelion distance and inclination are in phase. Figure 3.6 shows a rare example of such variations among short-period comets. Comet 96P was discovered by American amateur astronomer Donald Machholz in 1986. With a Tisserand parameter of only 1.94 in spite of a short orbital period, L_z is indeed small, and the perihelion distance spans a wide range. The comet is close to the borderline between the Jupiter Family and Halley Type comets, and hence its dynamical origin is not clear.

The mechanism behind such substantial variations of e and i was first explained in the framework of asteroid dynamics by Japanese astronomer Yoshihide Kozai (1962). The terms *Kozai resonance* for the 1:1 resonance of $\dot{\varpi}$ and $\dot{\Omega}$ and *Kozai cycle* for the coupled oscillation of e and i have become generally accepted, like the term *Kozai mechanism*. However, it has recently been recognized that the

Russian physicist Mikhail Lidov discovered the mechanism somewhat before Kozai in the different framework of satellite dynamics. Hence, the name Kozai is sometimes replaced by Lidov–Kozai.

The Lidov–Kozai mechanism is very important for comet evolution, as will be seen several times in the following chapters. Examples are the physical destruction of comets in sungrazing or Jupiter-impacting orbits attainable by this mechanism. We will now instead treat an issue of great importance for the arrival of the new comets from their distant source.

3.3.1. *Galactic tides*

Consider comets belonging to the Oort Cloud (see Sec. 1.4) at typical distances of more than 10 000 AU from the Sun. These are subject to a tidal force due to the whole Galaxy, which perturbs their orbital motions around the Sun. According to a simple axisymmetric model of the gravitational field of the Galaxy, this force has two components: one radial with respect to the Galactic center, and one perpendicular to the Galactic plane. Due to the offsets of classical Oort Cloud comets from the Sun in the radial or normal directions, this tidal force can be noticeable.

The tidal acceleration can be expressed in terms of the local density of the Galactic disk (ρ_o) and the kinematic parameters of Galactic differential rotation — the so-called Oort constants A and B. Using cartesian coordinates (x', y', z) centered on the Sun such that the unit vectors $\hat{\mathbf{x}}'$ and $\hat{\mathbf{y}}'$ point toward the Galactic anticenter and transversely along the local circular velocity in the Galactic plane, and $\hat{\mathbf{z}}$ is perpendicular to this plane, the equation of motion can be written

$$\begin{aligned}\ddot{\mathbf{r}} = &-\nabla U_\odot + (A-B)(3A+B)x'\hat{\mathbf{x}}' - (A-B)^2 y'\hat{\mathbf{y}}' \\ &- [4\pi G\rho_o - 2(B^2 - A^2)]z\hat{\mathbf{z}},\end{aligned} \tag{3.15}$$

where $U_\odot = -GM_\odot/r$ is the solar potential.

With modern estimates of $A = +13\,\mathrm{km/s/kpc}$ and $B = -13\,\mathrm{km/s/kpc}$ (Gunn *et al.* 1979) and $\rho_o = 0.1\,M_\odot/\mathrm{pc}^3$ (Holmberg and Flynn 2000), we realize that only the term involving ρ_o is

non-vanishing in the z component, and this term is almost ten times larger than the coefficients of the the x' and y' components. Thus the disk tide is much stronger than the radial tide.

By neglecting the latter, we hence get a first approximation to the long-term dynamical behavior of Oort Cloud comets by using the equation

$$\ddot{\mathbf{r}} = -\nabla \left\{ -\frac{GM_\odot}{r} + 2\pi G\rho_o z^2 \right\} \tag{3.16}$$

(Heisler and Tremaine 1986). Let us now note that the perturbing acceleration in Eq. (3.16) is equivalent to that of an infinite disk of matter in the Galactic plane with density ρ_o. This disk can be seen as an infinite superposition of annuli similar to the one representing Jupiter in the Kozai mechanism. Therefore, the result is similar, and the vertical Galactic tide of Eq. (3.16) causes perturbations that we may interpret as due to a Kozai cycle. Like in the case of the long-term jovian perturbations, we deal with an integrable dynamical system, and the solution can be described by analytical formulae.

Heisler and Tremaine (1986) developed such a theory. They averaged the comet's Hamiltonian function over the orbital period, so that the mean anomaly disappeared from the Hamiltonian equations of motion, thereby securing an energy integral. The conservation of L_z then made the system fully integrable, and the angular variable is the Galactic argument of perihelion ω_G. The resulting Lidov–Kozai cycle for the eccentricity is illustrated for two representative cases in Fig. 3.7. The plotted quantity $1 - e$ and the Galactic inclination i_G both vary with ω_G, in phase with each other.

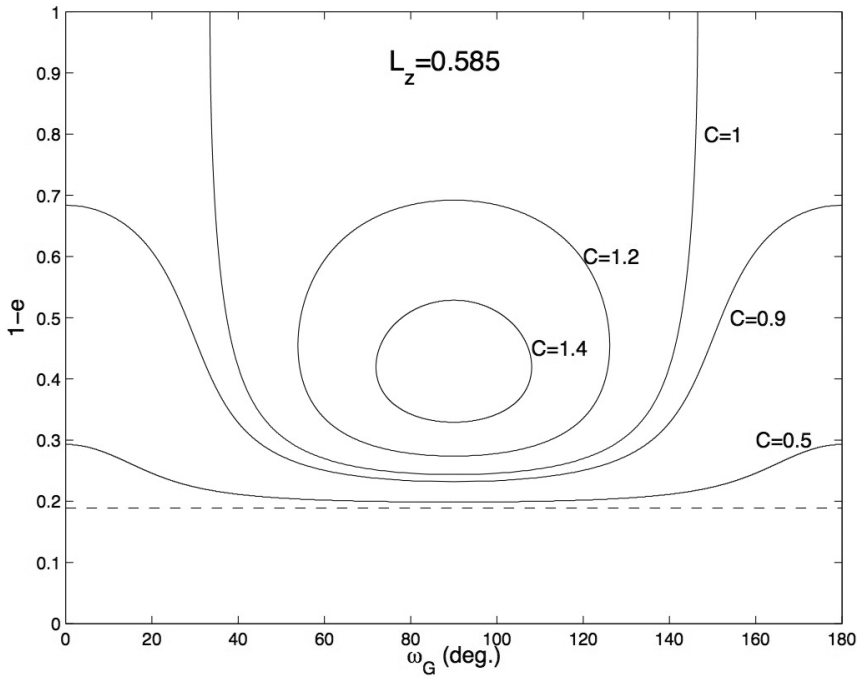
The integrals characterizing the curves are the energy, or Hamiltonian:

$$C = 1 - e^2 + 5e^2 \sin^2 i_G \sin^2 \omega_G, \tag{3.17}$$

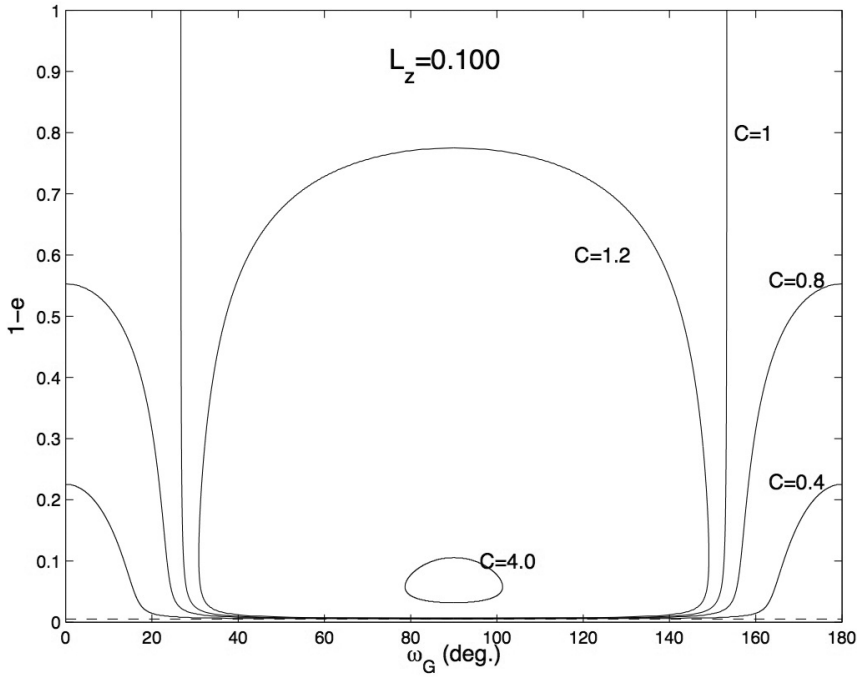
and the normalized, vertical angular momentum component:

$$L_z = \sqrt{1 - e^2} \cos i_G. \tag{3.18}$$

The separatrix between libration and circulation of ω_G corresponds to $C = 1$. As will be seen in Sec. 5.2, it is of interest to derive



(a)



(b)

Fig. 3.7. Variations of $1 - e$ with ω_G according to the Galactic disk tide theory by Heisler and Tremaine (1986). Panels (a) and (b) exhibit two values of the normalized, vertical angular momentum component $L_z = \sqrt{1 - e^2} \cos i_G$, and the curves show projections of the phase space trajectories for selected values of the Hamiltonian. The dashed lines correspond to $L = L_{\min} = L_z$. Modified from H. Rickman, "Cometary Dynamics," in Lecture Notes in Physics, Vol. 790 (2010), pp. 341–399, © Springer-Verlag Berlin Heidelberg 2010. With permission of Springer Nature.

an expression for the variation of the perihelion distance caused by the Galactic tide over one orbital revolution. For comets that are either observable or nearly so, the eccentricity is close to unity, which implies the approximation $L \propto \sqrt{q}$ for the absolute value of the angular momentum. Hence we have

$$\frac{dq}{dt} \propto \sqrt{q}\frac{dL}{dt}, \tag{3.19}$$

and the expression for the averaged Hamiltonian yields a formula for the rate of change of the angular momentum:

$$\left|\frac{dL}{dt}\right| \simeq 5\pi G\rho_o a^2 |\sin 2\beta_G|, \tag{3.20}$$

where β_G is the Galactic latitude of perihelion. For the change of perihelion distance during one orbital revolution, i.e., a time interval proportional to $a^{3/2}$, we get

$$|\Delta q| \propto \sqrt{q} a^{7/2} |\sin 2\beta_G|, \tag{3.21}$$

showing that the maximum changes occur for orbits with Galactic perihelion latitude equal to $\pm\pi/4$.

Let us finally note that the integrability of the equations of motion was obtained by the orbital averaging of the Hamiltonian and therefore rests upon the correctness of this procedure. A verification of this may be obtained by showing that the orbital period is much shorter than the period of ω_G libration. This is indeed the case for orbits with semi-major axis $a \sim 10\,000$ AU or less, but the libration period varies inversely with the orbital period, and for $a \gtrsim 40\,000$ AU the integrability breaks down. In addition, the radial component of the tidal acceleration plays an increasingly important role and may drive comets out of the Sun's Hill sphere in the Galaxy, whereby they are lost from the solar system.

3.4. Stellar Perturbations

Let us recall the meaning of the Galactic tide. This is the result of comets moving relatively far from the Sun in the gravitational field of the Galaxy. Here one considers the smooth potential of the entire stellar system similar to what is done when tracing stellar

orbits. But the analogy between cometary dynamics and Galactic dynamics reaches further. The smooth potential situation is a very good approximation most of the time, when the star is far from the potential wells of other stars, but encounters do occur so that the stellar orbits are deflected. The accumulated effect of such deflections over a very long time is called *relaxation*.

In the case of Oort Cloud comets, we also have to deal with stellar encounters. In principle, there are other objects with much larger mass, which the solar system may also encounter during its history. Giant molecular clouds (so-called GMCs) and stellar clusters are examples. But these are extremely rare, and in general they are neglected in analyses of Oort Cloud dynamics. Galactic field stars, on the other hand, are always present in the solar neighborhood, and they even penetrate into the Oort Cloud relatively frequently.

In terms of kinematics, the Sun is a normal member of the thin disk. Its excursions from a circular orbit in the Galactic mid-plane are small — both vertically and radially. This property is shared by a great majority of the nearby stars, so a local centroid of stellar motions can be defined as the average of their velocity vectors in an arbitrary reference frame. Relative to the circular motion, this centroid is somewhat lagging behind, since more stars are in the outer than the inner part of their Galactic orbits. Each star has a peculiar velocity with respect to the centroid, and so has the Sun as well. This amounts to about 20 km/s and is directed toward the *apex*, which is situated in the constellation of Hercules.

A crude estimate of the time scale for close encounters between the Sun and its neighbors can be derived using the mean stellar velocity (V_*) and the local number density of stars (n_*):

$$T_* = (\pi R^2 V_* n_*)^{-1}, \tag{3.22}$$

where encounters within a distance R are considered, and T_* is the average time between successive encounters. It is not enough to substitute the apex velocity for V_*, because the dispersion of the peculiar velocities of other stars is also very important.

Figure 3.8 shows the result of using Eq. (3.22) with $V_* = 30\,\mathrm{km/s}$ and $n_* = 0.1\,\mathrm{pc}^{-3}$. Let us estimate that the Oort Cloud extends to a

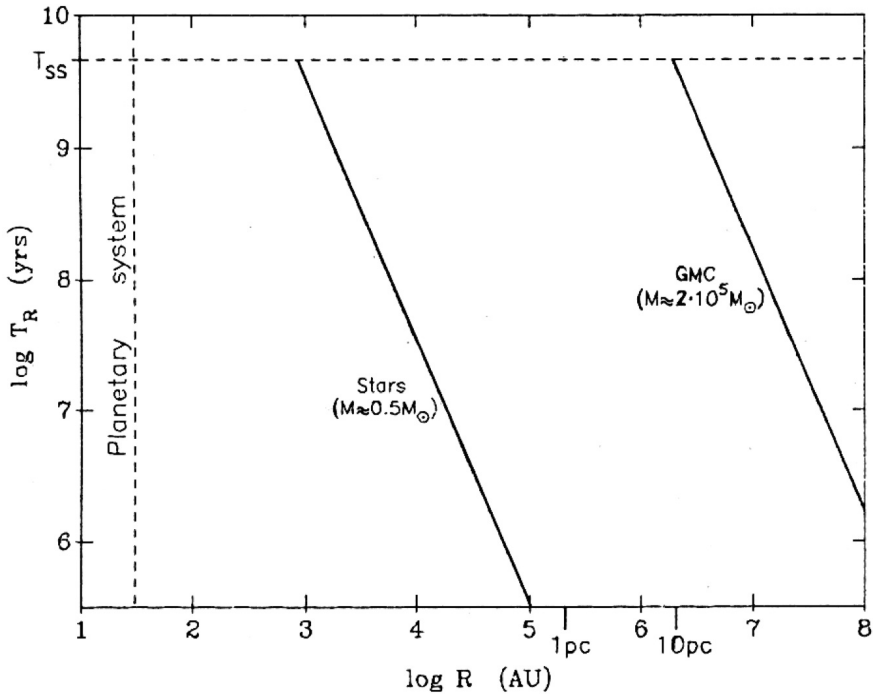


Fig. 3.8. The log of the average time interval between consecutive encounters between the Sun and stars of the current solar neighborhood (left line) or Giant Molecular Clouds (right line) versus the log of the miss distance. T_{SS} is the age of the solar system, and the vertical dashed line indicates the radius of the planetary system. Credit: H. Rickman, "Cometary Dynamics," in Lecture Notes in Physics, Vol. 790 (2010), pp. 341–399, © Springer-Verlag Berlin Heidelberg 2010. With permission of Springer Nature.

radius of about 50 000 AU. It then follows that the average interval between penetrating stellar encounters is about 1.7 million years. We easily realize that this is similar to — or even shorter than — the orbital periods of many Oort Cloud comets. Even though a typical stellar passage has a large effect only on a minority of these comets, it is clear that such effects have to be considered when discussing Oort Cloud dynamics.

Another piece of information contained in Fig. 3.8 is that the closest stellar encounter to be expected during the age of the solar system has a distance $R \sim 1\,000$ AU. This should not be taken literally, because the Sun has likely experienced neighborhoods quite different from the present one, when it was a young star, and this allows for a possibility of even closer encounters (see Sec. 6.2). On the other hand, no encounter with a GMC even close to the Oort Cloud radius is likely to have occurred.

We now turn to the dynamics of the stellar encounters. To simplify the problem, let us only consider the interactions between the Sun, one passing star and the comet. Since the comet is practically massless, this is a restricted three-body problem. This is usually a good approximation, since in most cases the star-star interactions are weak, and the effects of two stars on the Sun and the comet are thus additive, even if these stars should happen to encounter the Sun nearly simultaneously. However, we should be aware that stars often form binary or multiple systems. When the Sun encounters a binary star, the situation is benign in two limiting cases. Many binaries are very tight, in which case they may be approximated by a single point mass at their center of mass. Wide binaries form another frequent category, and then the two components can be treated as individual stars following two different but parallel trajectories. Between these two extremes there are in principle cases that require a more realistic treatment, but this is rarely done in practice.

The weakness of the interactions can be illustrated by using Eq. (3.8). Assume that a star approaches the Sun with a velocity of 30 km/s and an impact parameter of 50 000 AU. One can then derive a hyperbolic deflection angle of only eight arcseconds. Even if the star aims at only 10 000 AU from the Sun, the deflection amounts to less than one arcminute. It is thus common practice to approximate the stellar trajectories by straight lines when treating encounters with Galactic field stars.

It is thus possible to treat the problem in a heliocentric frame, where the star travels with constant speed along a straight line. Concerning the comet, the simplest treatment is to assume that it stays at rest during the passage of the star. The argument behind this assumption is that the speed of an Oort Cloud comet is limited by the speed of escape, which at a distance of 10 000 AU from the Sun amounts to 0.4 km/s. Since the speed of the star is ~ 100 times larger, it is natural to neglect the motion of the comet as a first approximation.

If one integrates the acceleration imparted to the Sun or the comet over time during the whole passage of the star, one obtains the resulting impulse or velocity change caused by the star. This

is a vector situated in the plane spanned by the accelerated object and the stellar trajectory. This has two components — one directed along the trajectory and the other perpendicular to it. The latter points from the object to the closest point on the trajectory. It is easily seen from the symmetry around the closest point that the accelerations caused by the star along its trajectory before and after crossing this point cancel out. Thus, the impulse vector reduces to the perpendicular component, which is readily shown to be

$$\mathbf{I} = \frac{2GM_*}{V_*}\,\frac{\hat{\mathbf{d}}}{d}, \tag{3.23}$$

where M_* is the mass of the star, V_* is its velocity, d is the distance from the object to the closest point on the trajectory, and $\hat{\mathbf{d}}$ is the unit vector along this direction.

The impulse imparted to the comet in the heliocentric frame is obtained as the difference between the cometary and solar impulses:

$$\Delta\mathbf{V}_c = \frac{2GM_*}{V_*}\left\{\frac{\hat{\mathbf{d}}_c}{d_c} - \frac{\hat{\mathbf{d}}_\odot}{d_\odot}\right\}. \tag{3.24}$$

By placing the comet at the desired place in its orbit and applying the impulse derived from Eq. (3.24) to the heliocentric velocity at this point, one can compute new orbital elements from the new velocity. This way of accounting for stellar perturbations is called the *Classical Impulse Approximation* (CIA), and it dates back to the earliest works where this problem was considered. The relevant geometry is illustrated in Fig. 3.9.

Nonetheless, it is worth considering the shortcomings of the CIA and ways to improve it. If, for instance, the comet moves in the direction of the vector $\mathbf{d}$ during the lapse of the stellar encounter, there will be a cometary impulse component in the direction of the stellar motion, which translates directly into the heliocentric impulse $\Delta\mathbf{V}_c$. Other motions of the comet will also have effects on the calculated impulse. Generally speaking, those effects grow in importance for comets in smaller orbits. For the inner core of the Oort Cloud (see Chap. 5), the CIA is not very accurate.

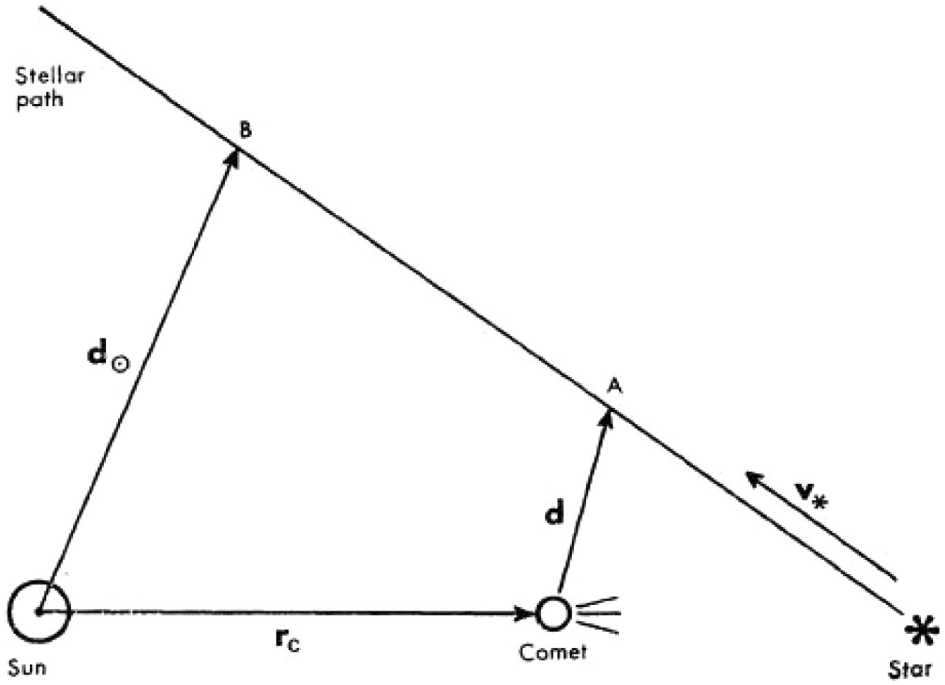


Fig. 3.9. The geometry of a stellar encounter as considered in the classical impulse approximation. The vectors $\mathbf{d}$ and $\mathbf{d}_\odot$ make right angles to the stellar trajectory at points A and B, respectively. © Publishing House of the Czechoslovak Academy of Sciences (1976).

Due to the approximations inherent in the treatment of stellar encounters as single, isolated events, there is no point in performing accurate numerical integrations. Improvements of the CIA have involved accounting for the actual, hyperbolic shape of the stellar trajectory (Dybczyński 1994) and splitting the stellar passage into finite segments and adding up the impulses received during each of these (Eggers and Woolfson 1996; Rickman *et al.* 2005). This sequential impulse approximation has been found to yield good results in terms of accuracy versus computing time for almost any kind of cometary orbit.

An undeniable advantage of the CIA is its ability to yield analytical results that allow us to estimate the statistical influence of stellar perturbations on comets at large. Since most stellar passages through the Oort Cloud do not involve particularly close encounters with any given comet or the Sun, they cannot be treated by neglecting any of the terms in Eq. (3.24). The only simplification that is generally valid to some extent is achieved by assuming the star to pass relatively far from the Sun-comet pair, as illustrated by Fig. 3.10.

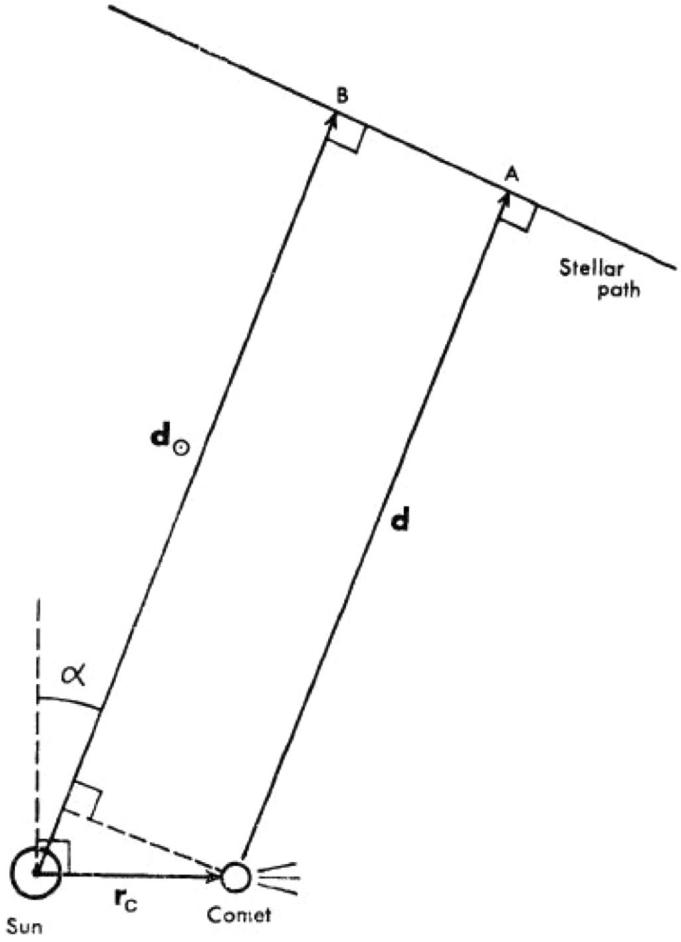


Fig. 3.10. The geometry of a distant stellar encounter leading to a tidal version of the classical impulse approximation. In this case, the differential impulse on the comet relative to the Sun is estimated. © Publishing House of the Czechoslovak Academy of Sciences (1976).

A tidal approximation to the heliocentric impulse then follows as:

$$\Delta \mathbf{V}_c \approx \frac{2GM_* r_c \sin\alpha}{V_* d_\odot^2} \cdot \hat{\mathbf{d}}_\odot, \tag{3.25}$$

where the symbols are as shown in the figure. Since the perturbation of the angular momentum is $\Delta \mathbf{L}_c = \mathbf{r}_c \times \Delta \mathbf{V}_c$, its absolute value is seen to be proportional to r_c^2. Thus, exposing comets with different semi-major axes a to the same set of stellar perturbations, we obtain $|\Delta L| \propto a^2$. Since the perturbation of the perihelion distance is approximately $\Delta q \propto \sqrt{q}\Delta L$, the expectance of $|\Delta q|$ is

$$\mathcal{E}(|\Delta q|) \propto \sqrt{q} \cdot a^2. \tag{3.26}$$

Now, consider a time interval short enough that no more than one stellar encounter can be expected. Its probability of occurrence is then proportional to the length of the interval. This is a fair approximation for the orbital period of an Oort Cloud comet, which is $P \propto a^{3/2}$. We hence get the expectance of the change in perihelion distance during one revolution as

$$\mathcal{E}(|\Delta q|_P) \propto \sqrt{q} \cdot a^{7/2}. \tag{3.27}$$

3.5. Energy Diffusion

In Sec. 3.3, we treated the secular evolution of comet orbits (in the absence of close encounters) under the assumption that the perturbing planet acts like a uniform, circular annulus of matter, thus causing no changes in the orbital energy of the comet. This is of course a highly idealized situation, which is useful to understand the secular changes in the angular momentum of the comets but which does not hold true in reality. In particular, the orbital energy of the comets does change, and we will now account for these variations.

The histogram in Fig. 3.11 from Rickman (2010) shows the distribution of changes of inverse semi-major axis experienced by observed long-period comets during their passages through the planetary system. The initial and final orbits in question are barycentric, i.e., referred to the joint center of mass of the Sun and the planets. The underlying comet sample is not biased in a dynamical sense, since the only criterion is that they were discovered and well enough observed to establish their orbits with good accuracy.

The distribution exhibits a narrow peak surrounded by wide tails. The latter are due to the few comets that had fairly close encounters with Jupiter and will not be discussed here. We estimate the half width at half maximum of the peak to be about $0.0006\,\mathrm{AU}^{-1}$. There is no obvious departure from symmetry around zero — positive and negative perturbations are about equally common.

For a theoretical background, we can write the equation of motion of a comet in the heliocentric frame as

$$\ddot{\mathbf{r}} = -\nabla U_\odot - \nabla R_p, \tag{3.28}$$

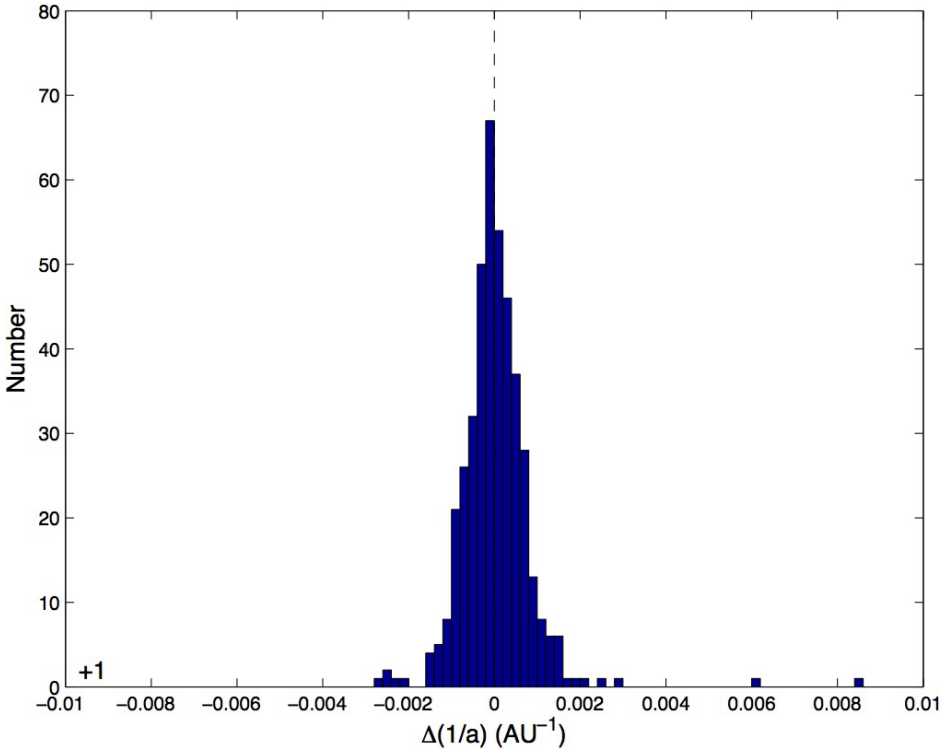


Fig. 3.11. Distribution of planetary perturbations of inverse, barycentric semi-major axis for 418 observed long-period comets, derived from data listed by Marsden and Williams (2005). Credit: H. Rickman, "Cometary Dynamics," in Lecture Notes in Physics, Vol. 790 (2010), pp. 341–399, © Springer-Verlag Berlin Heidelberg 2010. With permission of Springer Nature.

where $U_\odot = -GM_\odot/r$ is the solar gravitational potential and

$$R_p = -GM_p \left\{ \frac{1}{|\mathbf{r} - \mathbf{r}_p|} - \frac{\mathbf{r}_p \cdot \mathbf{r}}{r_p^3} \right\} \tag{3.29}$$

is the so-called perturbing function, which looks like a potential in Eq. (3.28) but varies with time due to the motions of the comet and the planet (radius vectors $\mathbf{r}$ and $\mathbf{r}_p$), as seen from Eq. (3.29). It has to be accurately monitored when solving the equation of motion. Of course, when several planets are involved, R_p is replaced by the sum of all such expressions for the different planets.

We are now interested in the rate of change of the inverse semi-major axis of the comet. For this purpose, we note that the orbital energy is given by

$$E = \frac{1}{2}\dot{\mathbf{r}}^2 + U_\odot = -\frac{GM_\odot}{2a}, \tag{3.30}$$

where a is the semi-major axis. By taking the time derivative of Eq. (3.30) we get

$$GM_\odot \frac{d}{dt}\left(\frac{1}{2a}\right) = -\dot{\mathbf{r}} \cdot \ddot{\mathbf{r}} - \dot{\mathbf{r}} \cdot \nabla U_\odot, \tag{3.31}$$

and by taking the scalar product of Eq. (3.28) with $\dot{\mathbf{r}}$, we recognize that

$$GM_\odot \frac{d}{dt}\left(\frac{1}{2a}\right) = -\dot{\mathbf{r}} \cdot GM_p \nabla \left\{ \frac{1}{|\mathbf{r} - \mathbf{r}_p|} - \frac{\mathbf{r}_p \cdot \mathbf{r}}{r_p^3} \right\}. \tag{3.32}$$

Introducing $m_p = M_p/M_\odot$ and the dimensionless position vectors $\mathbf{s} = \mathbf{r}/a_p$ and $\mathbf{s}_p = \mathbf{r}_p/a_p$ with the planetary semi-major axis a_p as unit of length, we finally get

$$\frac{d}{dt}\left(\frac{1}{a}\right) = \frac{2m_p}{a_p} \dot{\mathbf{s}} \cdot \left\{ \frac{(\mathbf{s}_p - \mathbf{s})}{|\mathbf{s}_p - \mathbf{s}|^3} + \frac{\mathbf{s}_p}{s_p^3} \right\}. \tag{3.33}$$

The contribution by any planet to the time derivative of $1/a$ is thus proportional to m_p/a_p times an expression depending on the geometrical configuration and the velocity of the comet. In the common situation, when the comet is quite far from every planet, it is clear that the jovian contribution dominates over those of the other planets. Thus, the width of the central peak in the perturbation distribution of Fig. 3.11 is essentially set by the jovian perturbations.

A very accurate determination of such a perturbation requires a numerical integration of the equations of motion, yielding the functions $\mathbf{s}(t)$ for the comet and $\mathbf{s}_p(t)$ for all the planets. However, an approximate solution can be obtained by considering the unperturbed, Keplerian orbits and the corresponding, analytically derived position vectors. This method must be reserved for perturbations that change the comet orbit only insignificantly so that the action of each planet is independent of those of other planets.

Now, consider for instance the typical jovian perturbations that build up the central peak of the distribution. These can be positive or negative, and larger or smaller, the outcome depending on the time

sequence of geometrical configurations exhibited by the comet and Jupiter. Of course, the different comets that contribute to Fig. 3.11 are independent of each other, and their respective geometrical patterns are uncorrelated. But is this the case also for the consecutive passages of the same comet, as it returns to perihelion on successive orbits? If so, the sequence of perturbations would be like a set of random drawings from a relevant parent distribution. Alternatively, will the comet retain a memory of its previous perihelion passage, as it passes perihelion again, so that the geometries — and thus the perturbations — are in fact correlated? If so, there would be some regularity or predictability in the sequence of perturbations.

If we consider comets with very long periods — say, tens of thousands of years or more — there is no regularity in the sequence, and the dynamical evolution appears like a random walk along the energy axis. The reason why such comets behave this way is that the geometrical pattern followed by a comet during a given perihelion passage is extremely sensitive to the way the comet was perturbed during the previous perihelion passage. Even an insignificant change in this perturbation may change the timing of the following perihelion passage by several years, and this is enough to cause a major change in the perturbation exerted by Jupiter.

The resulting unpredictability is the signum of *dynamical chaos*, which means that nearby orbits diverge exponentially on a time scale called the *Lyapunov time*. The orbital evolution of long-period comets is hence chaotic, and the random walk they experience along the energy axis is an expression of this chaos. As long as the steps are small, representing the central peak of the probability distribution, a relevant mathematical description is offered by diffusion theory. This has been frequently used in early works but suffers from important limitations due to the existence of the wide tails surrounding the central peak. These tails do not extend to infinity, and the very large perturbations are extremely rare, but given enough time they may dominate the total effect.

When comets with shorter periods are considered, a limit may be reached, beyond which the random walk picture is no longer relevant. A small difference in the perturbation at one perihelion passage

will no longer lead to drastic changes in the following perturbation. Nearby orbits will still diverge rapidly, but the chaos is gone. In fact, resonant behavior starts to appear so that the energy may oscillate, as the comet orbit librates around a commensurability with Jupiter's mean motion. This is known to occur for some Halley Type comets (Carusi *et al.* 1987). The mentioned limit is not sharp but is usually considered to correspond to a semi-major axis of ~ 100 AU.

Let us finally mention a somewhat different type of chaos, which occurs for Jupiter Family comets and the related Centaurs. This has its roots in close encounters — in particular, with Jupiter. From the above description of close encounter dynamics, it is clear that the outcome of a hyperbolic deflection is sensitive to the initial conditions. Hence, the difference between two neighboring heliocentric orbits gets blown up considerably during a close encounter, and if a new encounter occurs only a few decades or a century later, the perturbations may differ drastically. Over longer times, a practical identity of two initial orbits may become totally unrecognizable. An example of this is shown in Fig. 3.12.

The difference between the chaos expressed by the random walk of long-period comet energies and the semi-major axis evolutions plotted in Fig. 3.12 for 95P/Chiron is that the latter chaos is intermittent, arising from short episodes (i.e., close encounters) between which the evolution appears quite regular. Nevertheless, chaos is pervasive in the long-term orbital evolutions of Jupiter Family comets and Centaurs in general, and the reason is the close encounters. Between those, the objects in question appear stable or show regular oscillations, including librations around mean motion resonance with Jupiter.

In the latter case, the commensurability of the orbital periods can stabilize the comet orbit by offering protection against close encounters. Consider for instance the 2:1 resonance between a comet and Jupiter. Typically, the aphelion of the comet orbit is close to Jupiter's orbit, and the perihelion is much closer to the Sun. If Jupiter is 90° away from the comet's aphelion direction, when aphelion passage occurs, there is no close encounter. In exact resonance, Jupiter has moved around half its orbit at the time of the next

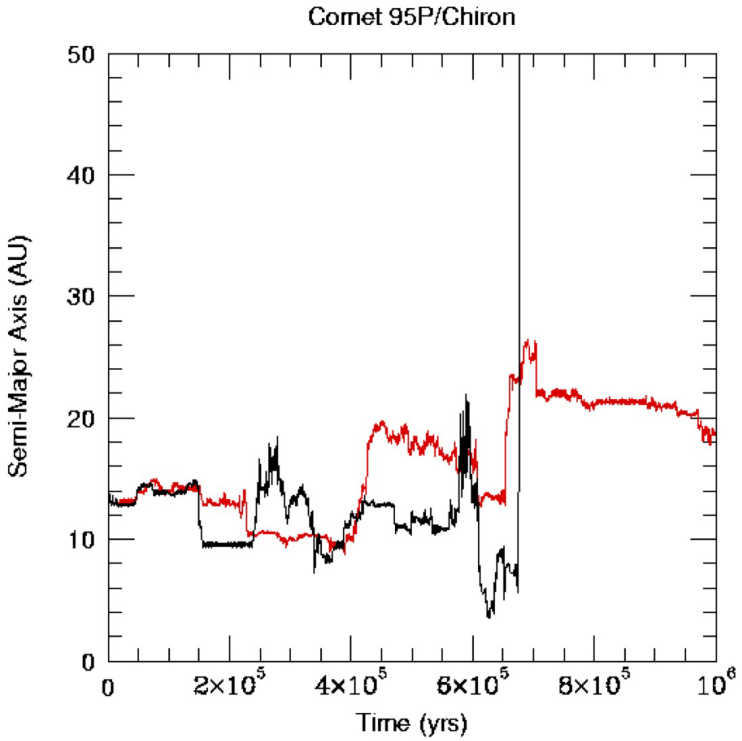


Fig. 3.12. Orbital evolution of the Centaur 95P/Chiron, obtained from accurate numerical integrations. The semi-major axis is shown as a function of time from the start of the integration. The red and black curves represent two variants with a minute difference in the initial conditions. Due to sequences of close encounters with giant planets, the two orbits diverge considerably within $\sim 10^5$ years. Hence, the two objects lose all memory of their initial vicinity, and the dynamics is clearly chaotic. Courtesy H. F. Levison. Credit: H. Rickman, "Cometary Dynamics," in Lecture Notes in Physics, Vol. 790 (2010), pp. 341–399, © Springer-Verlag Berlin Heidelberg 2010. With permission of Springer Nature.

aphelion passage and is therefore 90° away again. The same situation is bound to repeat over and over. However, if the resonance is not exact, the angle between the radius vectors of the comet and Jupiter at the comet's aphelion passage will change, so that one of the two consecutive aphelion passages gets more and more dangerous as the angle decreases. It can be shown that Jupiter's gravity tends to counteract this evolution, so that the decrease can be halted and the angle starts increasing again. Hence the angle may oscillate or "librate" around 90°, and close encounters can be avoided as long as this libration persists. However, in reality there are always small interferences by other planets, and a librating comet is bound to have a large libration amplitude to begin with. Thus, the librations are always short-lived, being broken by close encounters.

3.6. General Road Map

We have described the elements of comet dynamics in terms of the dynamical processes that govern the changes of their orbits. The perturbing agents are the planets, the passing stars, and the Galaxy as a whole. This has demonstrated that comet orbits are generally unstable and subject to major changes over long enough time intervals. Let us now sketch the patterns of orbital evolution that arise from the perturbations suffered by the comets.

Observable comets are ephemeral. Being Jupiter-crossing, their orbits are unstable and in addition, the observable comets may end their lives by physical effects, as will be discussed in Chap. 4. Due to the transfer routes provided by comet dynamics, the observable population is replenished and may be kept in a steady state. The source for these transfers is present in distant parts of the solar system in the form of giant reservoirs of icy bodies, which turn into comet nuclei as they enter into observable orbits.

Our description of the transfer routes had better start from the distant sources. Let us first consider the most distant of all — the Oort Cloud. This includes comets with a wide range of semi-major axes. While they all are subject to external perturbations due to Galactic tides and stellar encounters, the effects are the strongest for the most distant comets. The orbital energy is not immune to these perturbations, but the most profound changes are suffered by the angular momentum. There are hence large-scale changes of the perihelion distance, which may bring comets all the way between nearly circular orbits and nearly parabolic ones with perihelia in the observable region. This is the origin of the so-called *new comets*, which have the largest orbital energies (i.e., negative but closest to zero) among all observed comets.

Oort Cloud comets with smaller semi-major axes experience the same kind of external perturbations but on longer time scales. Their routes into observable orbits, to be described in Chap. 5, differ from those of the more distant comets but nonetheless end up in the same category of observed, new comets. In the current Galactic environment there is a limit to the semi-major axis, below which the perturbation time scale grows larger than the age of the solar

system. Comets in those orbits would be isolated from the rest of the solar system and would never become observable. If the Oort Cloud does include such fossilized comets, these must have found their way into their current abode at a time, when the solar system was placed in a different environment. We shall return to such possibilities in Chap. 6, but for now there is no reason to discuss comets situated beyond the current transfer routes.

Let us instead return to the new comets. The constituent comets of the Oort Cloud often have semi-major axes in excess of 10 000 AU, and thus the new comets typically have inverse semi-major axes between zero and 0.0001 AU^{-1}. Since the typical perturbations of $1/a$ according to Fig. 3.11 are several times larger than this, it follows that the new comets are very often expelled from the solar system on hyperbolic orbits. The subsequent Galactic and stellar perturbations will rarely protect them from expulsion, and thus the solar system feeds comets into interstellar space as long as the flux of comets from the Oort Cloud persists.

However, at least as many new comets experience the opposite perturbations, thus becoming more tightly bound in orbits with smaller semi-major axes. A certain fraction will be lucky enough to be captured more tightly than the average, and these stand a very good chance of experiencing many more perturbations on subsequent returns. The resulting evolution is a random walk in orbital energy as described above. A salient feature of this evolution is the presence of an absorbing wall at zero energy, because there is no return from expulsion from the solar system. Increasing the number of perihelion passages (N), more and more comets are expelled, and the number of survivors decreases as $1/\sqrt{N}$. But these include an increasing fraction of comets that arrive into orbits, where the periods are too short for the random walk picture to apply. Even Halley Type comets might result from such an evolution.

In Fig. 3.13 we see a comprehensive outline of the cometary transfer routes in the solar system. The ones just described are included together with some that remain to be discussed. The latter involve a different category of comets. In fact, we are dealing with the two classes defined in Chap. 1.4: nearly isotropic comets and ecliptic

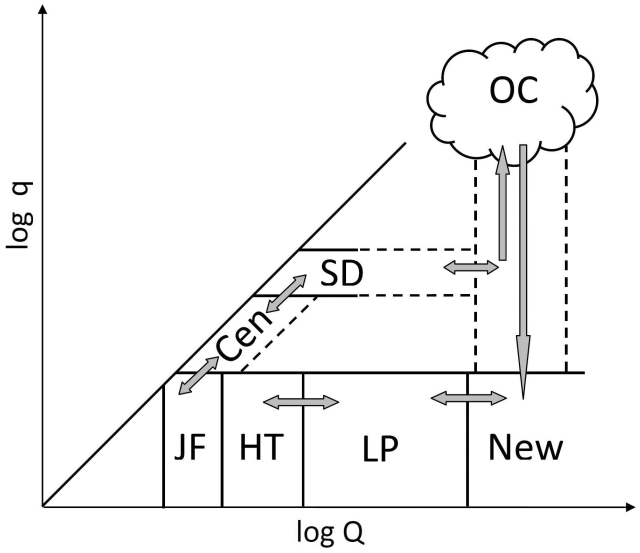


Fig. 3.13. Illustration of the main transfer routes of comets in the solar system using a generic diagram with the logs of the aphelion and perihelion distances on the axes. The routes are shown by shaded arrows. OC is the Oort Cloud; SD is the Scattered Disk; Cen means Centaurs; New stands for new comets; LP means long-period comets; HT means Halley Type comets; and JF is the Jupiter Family. Courtesy T. Wiśniowski.

comets. The Oort Cloud is the source of the nearly isotropic comets, while the ecliptic comets come from the flattened trans-neptunian populations — in particular, the *Scattered Disk*, which will be the subject of extensive discussions in Chaps. 5, 6 and 7. The observable component of the ecliptic comet population is the Jupiter Family. An indication of how comets are transferred into the Jupiter Family was obtained in Sec. 3.1 from Fig. 3.1, showing evolutionary curves for constant Tisserand parameter.

These curves connect the Jupiter Family with a part of the Centaur population, where some Chiron-type comets are found. In fact, as will be seen in Chap. 5, the Centaurs have a similar connection to the Scattered Disk and may be seen as a station on the way from the Scattered Disk into the Jupiter Family.

Finally, is there also a connection between the Scattered Disk and the Oort Cloud? The answer is yes. It is also easy to see the main direction of this transfer, namely, from the Scattered Disk into the Oort Cloud. Comets must have formed much closer to the Sun than both the Oort Cloud and the outer part of the Scattered Disk. It is then natural for the new-born comets to move

on low-inclination orbits, forming an initially very flat population. The Scattered Disk has some resemblance to this, even though the difference is substantial. As we shall see in Sec. 6.3.1, this disk is thought to have been created out of a primordial reservoir of comets. It then acted as a source, from which the Oort Cloud was built.

The dynamics of this process has some similarity to the above-described energy diffusion of long-period comets. However, the scattered disk objects have their perihelia near or slightly beyond the orbit of Neptune, and they experience gravitational scattering due to Neptune rather than Jupiter. Close encounters are once again rare, but in the long run they dominate the scattering. The time scale is measured in hundreds of millions of years. As a result, more and more disk objects are placed into orbits that reach out far enough for external perturbations to perturb their angular momenta, thus lifting the perihelia far away from all planets. Due to those perturbations and the following stellar encounters during billions of years, the orbits get fully randomized, losing the memory of an ecliptic origin. As we shall see in Sec. 6.3.2, this holds at least for the outer part of the Oort Cloud, while the inner parts still maintain a preference for low inclinations.

Of course, there is nothing to prevent Oort Cloud comets on their way toward observable orbits to have this process interrupted by close encounters with Neptune or any other giant planet. If they hence get decoupled from the external perturbations, they may get captured into scattered disk or Centaur orbits and eventually penetrate into Jupiter Family or Halley Type orbits. It is important to realize that all dynamical processes in a dissipation-free system are time reversible, but the question is which processes dominate in the real solar system. This question is often controversial, and sometimes the answers are only preliminary, as we shall see in later chapters.