The Kuiper Belt and the Primordial Evolution of the Solar System

A. Morbidelli
Observatoire de la Côte d'Azur

M. E. Brown
California Institute of Technology

We discuss the structure of the Kuiper belt as it can be inferred from the first decade of observations. In particular, we focus on its most intriguing properties — the mass deficit, the inclination distribution, and the apparent existence of an outer edge and a correlation among inclinations, colors, and sizes — which clearly show that the belt has lost the pristine structure of a dynamically cold protoplanetary disk. Understanding how the Kuiper belt acquired its present structure will provide insight into the formation of the outer planetary system and its early evolution. We critically review the scenarios that have been proposed so far for the primordial sculpting of the belt. None of them can explain in a single model all the observed properties; the real history of the Kuiper belt probably requires a combination of some of the proposed mechanisms.

1. INTRODUCTION

When Edgeworth and Kuiper conjectured the existence of a belt of small bodies beyond Neptune — now known as the Kuiper belt — they certainly were imagining a disk of planetesimals that preserved the pristine conditions of the protoplanetary disk. However, since the first discoveries of transneptunian objects, astronomers have realized that this picture is not correct: The disk has been affected by a number of processes that have altered its original structure. The Kuiper belt may thus provide us with many clues to understand what happened in the outer solar system during the primordial ages. Potentially, the Kuiper belt might teach us more about the formation of the giant planets than the planets themselves. And, as in a domino effect, a better knowledge of giant-planet formation would inevitably boost our understanding of the subsequent formation of the terrestrial planets. Consequently, Kuiper belt research is now considered a top priority in modern planetary science.

A decade after the discovery of 1992 QB_1 (*Jewitt and Luu,* 1993), we now know 770 transneptunian objects (semimajor axis $a > 30$ AU) (all numbers are as of March 3, 2003). Of these, 362 have been observed during at least two oppositions, and 239 during at least three oppositions. Observations at two and three oppositions are necessary for the Minor Planet Center to compute the objects' orbital elements with moderate and good accuracy respectively. Therefore, the transneptunian population is gradually taking shape, and we can start to seriously examine the Kuiper belt structure and learn what it has to teach us. We should not forget, however, that our view of the transneptunian population is still partial and is strongly biased by a number of factors, some of which cannot be easily modeled.

A primary goal of this chapter is to present the orbital structure of the Kuiper belt as it stands based on the current observations. We start in section 2 by presenting the various subclasses that constitute the transneptunian population. Then in section 3 we describe some striking properties of the population, such as its mass deficit, inclination excitation, radial extent and a puzzling correlation between orbital elements and physical properties. In section 4 we finally review the models that have been proposed so far on the primordial sculpting of the Kuiper belt. Some of these models date from the very beginning of Kuiper belt science, when only a handful of objects were known, and have been at least partially invalidated by the new data. Paradoxically, however, as the data increase in number and quality, it becomes increasingly difficult to explain all the properties of the Kuiper belt in the framework of a single scenario. The conclusions are in section 5.

2. TRANSNEPTUNIAN POPULATIONS

The transneptunian population is "traditionally" subdivided in two subpopulations: the scattered disk and the Kuiper belt. The definition of these subpopulations is not uniform, as the Minor Planet Center and various authors often use slightly different criteria. Here we propose and discuss a categorization based on the dynamics of the objects and their relevance to the reconstruction of the primordial evolution of the outer solar system.

In principle, one would like to call the Kuiper belt the population of objects that, even if characterized by chaotic dynamics, do not suffer close encounters with Neptune and thus do not undergo macroscopic migration in semimajor axis. Conversely, the bodies that are transported in semi-

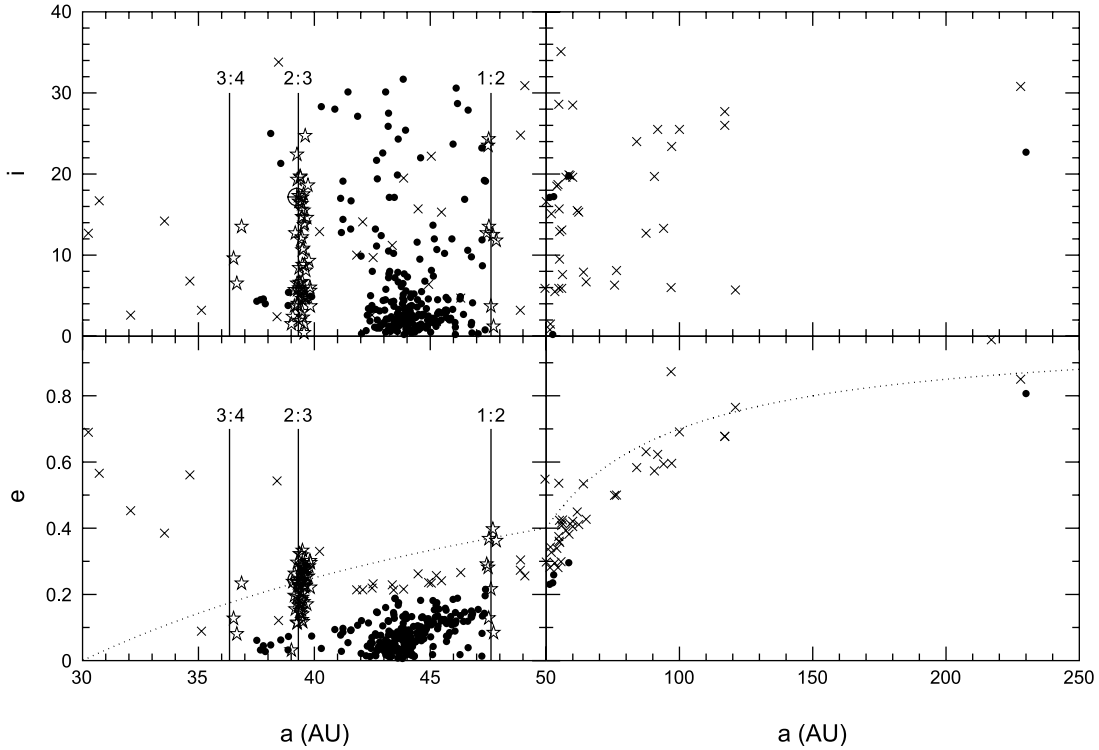

Fig. 1. The orbital distribution of multiopposition transneptunian bodies, as of March 3, 2003. Scattered disk bodies are represented as a cross, classical Kuiper belt bodies as dots, and resonant bodies as stars. In the absence of long-term numerical integrations of the evolution of all the objects and because of the uncertainties in the orbital elements, it is possible that some bodies could have been misclassified. The figure should thus be considered as an indicative representation of the various subgroups that compose the transneptunian population. The dotted curve denotes q = 30 AU. The vertical solid lines mark the locations of the 3:4, 2:3, and 1:2 mean-motion resonances with Neptune. The orbit of Pluto is represented by a crossed circle.

major axis by close and distant encounters with Neptune would constitute the scattered disk. The problem with precisely dividing the transneptunian population into Kuiper belt or scattered disk is related to timescale. On what timescale should we see semimajor axis migration resulting in the classification of an object in the scattered disk? The question is relevant, because it is possible for bodies trapped in resonances to significantly change their perihelion distance and pass from a scattering phase to a nonscattering phase (and vice versa) numerous times over the age of the solar system.

For this reason, we prefer to link the definition of the scattered disk to its formation mechanism. We refer to the scattered disk as the region of orbital space that can be visited by bodies that have encountered Neptune within a Hill's radius at least once during the age of the solar system, assuming no substantial modification of the planetary orbits. We then refer to the Kuiper belt as the complement of the scattered disk in the a > 30 AU region.

To categorize the observed transneptunian bodies into scattered disk and Kuiper belt, we refer to previous work on the dynamics of transneptunian bodies in the framework of the current architecture of the planetary system. For the a < 50 AU region, we use the results by *Duncan et al.* (1995) and *Kuchner et al.* (2002), who numerically mapped the regions of the (a, e, i) space with 32 < a < 50 AU that can lead to a Neptune-encountering orbit within 4 G.y. Because dynamics are reversible, these are also the regions that can be visited by a body after having encountered the planet. Therefore, according to our definition, they constitute the scattered disk. For the a > 50 AU region, we use the results by *Levison and Duncan* (1997) and *Duncan and Levison* (1997), who followed for a time span of another 4 G.y. the evolutions of the particles that encountered Neptune in *Duncan et al.* (1995). Despite the fact that the initial conditions did not cover all possible configurations, we can reasonably assume that these integrations cumulatively show the regions of the orbital space that can be possibly visited by bodies transported to a > 50 AU by Neptune encounters. Again, according to our definition, these regions constitute the scattered disk.

In Fig. 1 we show the (a, e, i) distribution of the transneptunian bodies that have been observed during at least two oppositions. The bodies that belong to the scattered disk according to our criterion are represented as crosses, while Kuiper belt bodies are represented by dots and stars (see explanation of the difference below).

We believe that our definition of scattered disk and Kuiper belt is meaningful for what concerns the major goal of Kuiper belt science, i.e., to reconstruct the primordial

evolution of the outer solar system. In fact, all bodies in the solar system must have been formed on orbits with very small eccentricities and inclinations, typical of an accretion disk. In the framework of the current architecture of the solar system, the current orbits of scattered disk bodies might have started with quasicircular orbits in Neptune's zone by pure dynamical evolution. Therefore, they do not provide any relevant clue to uncover the primordial architecture. The opposite is true for the orbits of the Kuiper belt objects with nonnegligible eccentricity and/or inclination. Their existence reveals that some excitation mechanism that is no longer at work occurred in the past (see section 4).

In this respect, the existence of Kuiper belt bodies with a > 50 AU on highly eccentric orbits is particularly important (five objects in Fig. 1, although our classification is uncertain for the reasons explained in the figure caption). Among them, 2000 CR_{105} (a = 230 AU, perihelion distance q = 44.17 AU, and inclination i = 22.7°) is a challenge by itself concerning the explanation of its origin. We call these objects extended scattered disk objects for two reasons: (1) they do not belong to the scattered disk according to our definition but are very close to its boundary and (2) a body of ~300 km like 2000 CR_{105} presumably formed much closer to the Sun, where the accretion timescale was sufficiently short (*Stern,* 1996), implying that it has been subsequently transported in semimajor axis until reaching its current location. This hypothesis suggests that in the past the true scattered disk extended well beyond its present boundary in perihelion distance. Given that the observational biases rapidly become more severe with increasing perihelion distance and semimajor axis, the currently known extended scattered disk objects may be the tip of the iceberg, e.g., the emerging representatives of a conspicuous population, possibly outnumbering the scattered disk population (*Gladman et al.,* 2002).

In addition to the extended scattered disk, we distinguish two other subpopulations of the Kuiper belt. We refer to the Kuiper belt bodies that are located in some major mean-motion resonance with Neptune [essentially the 3:4, 2:3, and 1:2 resonances (star symbols in Fig. 1) but also the 2:5 resonance (see *Chiang et al.,* 2003)] as the resonant population. It is well known that mean-motion resonances offer a protection mechanism against close encounters with the resonant planet (*Cohen and Hubbard,* 1965). For this reason, the resonant population — which, as part of the Kuiper belt, by definition must not encounter Neptune within the age of the solar system — can have perihelion distances much smaller than the other Kuiper belt objects, and even Neptune-crossing orbits (q < 30 AU) as in the case of Pluto. The bodies in the 2:3 resonance are often called Plutinos because of the analogy of their orbit with that of Pluto. We call the collection of Kuiper belt objects with a < 50 AU that are not in any notable resonant configuration the classical belt. Because they are not protected from close encounters with Neptune by any resonance, the stability criterion confines them to the region with small to moderate eccentricity, typically on orbits with q > 35 AU. The adjective "classical" is justified because, among all subpopulations, this is the one whose orbital properties are the most similar to those expected for the Kuiper belt prior to the first discoveries. We note, however, that the classical population is not that "classical." Although moderate, the eccentricities are larger than those that should characterize a protoplanetary disk. Moreover, several bodies have very large inclinations (see section 3.2). Finally, the total mass is only a small fraction of the expected pristine mass in that region (section 3.1). All these elements indicate that the classical belt has also been affected by some primordial excitation and depletion mechanism(s).

3. STRUCTURE OF THE KUIPER BELT

3.1. Missing Mass of the Kuiper Belt

The original argument followed by *Kuiper* (1951) to conjecture the existence of a band of small planetesimals beyond Neptune was related to the mass distribution in the outer solar system. The minimum mass solar nebula inferred from the total planetary mass (plus lost volatiles) smoothly declines from the orbit of Jupiter until the orbit of Neptune (see Fig. 2); why should it abruptly drop beyond the last planet? However, while Kuiper's conjecture on the existence of a transneptunian belt is correct, the total mass in the 30–50-AU range inferred from observations is two orders of magnitude smaller than the one he expected.

Kuiper's argument is not the only indication that the mass of the primordial Kuiper belt had to be significantly larger.

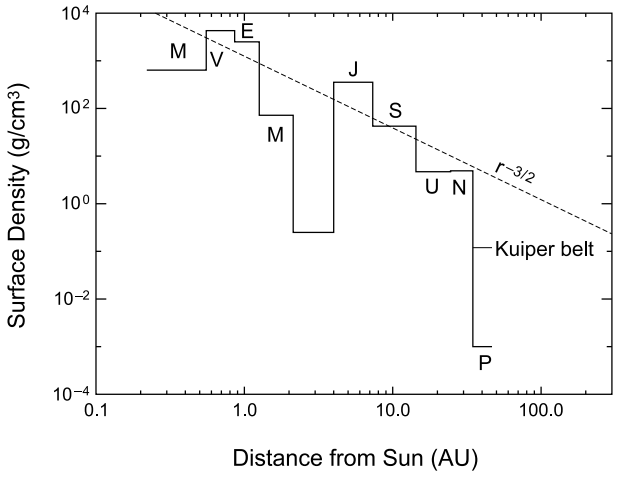

Fig. 2. The mass distribution of the solar nebula inferred from the masses of the planets augmented by the mass needed to bring the observed material to solar composition (data from *Lewis,* 1995). The surface density in the Kuiper belt has been computed assuming a current mass of ~0.1 $M_\oplus$ (*Jewitt et al.,* 1996; *Chiang and Brown,* 1999; *Trujillo et al.,* 2001; *Gladman et al.,* 2001) in the 42–48-AU annulus, and scaling the result by a factor of 70 in order to account for the inferred primordial local ratio between volatiles and solids. The estimate of the total mass in the Kuiper belt overwhelms that of Pluto, but still does not bring the mass to the extrapolation of the ~$r^{-3/2}$ line.

Further evidence for a massive primordial Kuiper belt was uncovered by *Stern* (1995), who found that the objects currently in the Kuiper belt were incapable of having formed in the present environment: Collisions are sufficiently infrequent that 100-km objects cannot be built by pairwise accretion of the current population over the age of the solar system. Moreover, owing to the large eccentricities and inclinations of Kuiper belt objects — and consequently to their high encounter velocities — collisions that do occur tend to be erosive rather than accretional, making bodies smaller rather than larger. Stern suggested that the resolution of this dilemma is that the primordial Kuiper belt was both more massive and dynamically colder, so that more collisions occurred, and they were gentler and therefore generally accretional.

Following this idea, detailed modeling of accretion in a massive primordial Kuiper belt was performed by *Stern* (1996), *Stern and Colwell* (1997a,b), and *Kenyon and Luu* (1998, 1999a,b). While each model includes different aspects of the relevant physics of accretion, fragmentation, and velocity evolution, the basic results are in approximate agreement. First, with ~10 $M_\oplus$ (Earth mass) or more of solid material in an annulus from about 35 to 50 AU on very low eccentricity orbits ($e \leq 0.001$), all models naturally produce a few objects on the order of the size of Pluto and approximately the right number of ~100-km objects, on a timescale ranging from several 10^7 to several 10^8 yr. The models suggest that the majority of mass in the disk was in bodies approximately 10 km and smaller. The accretion stopped when the formation of Neptune or other dynamical phenomena (see section 4) began to induce eccentricities and inclinations in the population high enough to move the collisional evolution from the accretional to the erosive regime (*Stern,* 1996). A massive and dynamically cold primordial Kuiper belt is also required by the models that attempt to explain the formation of the observed numerous binary Kuiper belt objects (*Goldreich et al.,* 2002; *Weidenshilling,* 2002).

Therefore, the general formation picture of an initial massive Kuiper belt appears secure. However, a fundamental question remains to be addressed: How did the initial mass disappear? Collisions can grind bodies down to dust particles, which are subsequently transported away from the belt by radiation pressure and/or Poynting Robertson drag, causing a net mass loss. The major works on the collisional erosion of a massive primordial belt have been done by *Stern and Colwell* (1997b) and *Davis and Farinella* (1997, 1998), achieving similar conclusions (for a review, see *Farinella et al.,* 2000). As long as the planetesimal disk was characterized by small eccentricities and inclinations, the collisional activity could only moderately reduce the mass of the belt. However, when the eccentricities and inclinations became comparable to those currently observed, bodies smaller than 50–100 km in diameter could be effectively destroyed. The total amount of mass loss depends on the primordial size distribution. To reduce the total mass from 30 $M_\oplus$ to a fraction of an Earth mass, the primordial size distribution had to be steep enough that essentially all the mass was carried by these small bodies, while the number of bodies larger than ~100 km had to be basically equal to the present number. Although this outcome seems consistent with the suggestions of the accretional models, there is circumstantial (but nonetheless compelling) evidence suggesting the primordial existence of a much larger number of large bodies (*Stern,* 1991). The creation of Pluto-Charon likely required the impact of two approximately similar-sized bodies that would be the two largest currently known bodies in the Kuiper belt. The probability that the two largest bodies in the belt would collide and create Pluto-Charon is vanishingly small, arguing that many bodies of this size must have been present and subsequently disappeared. Similarly, the existence of Triton and the large obliquity of Neptune are best explained by the existence at one time of many large bodies being scattered through the Neptune system. The elimination of these large bodies (if they existed in the Kuiper belt) could not be due to the collisional activity, but requires a dynamical explanation.

Another constraint against the collisional grinding scenario is provided by the preservation of the binary Kuiper belt objects. The Kuiper belt binaries have large separations, so it can be easily computed that the impact on the satellite of a projectile 100 times less massive with a speed of 1 km/s would give the former an impulse velocity sufficient to escape to an unbound orbit. If the collisional activity was strong enough to cause an effective reduction of the overall mass of the Kuiper belt, these kinds of collisions had to be extremely common, so we would not expect a significant fraction of widely separated binary objects in the current remaining population (*Petit and Mousis,* 2003.)

Understanding the ultimate fate of the 99% of the initial Kuiper belt mass that is no longer in the Kuiper belt is the first step in reconstructing the history of the outer solar system.

3.2. Excitation of the Kuiper Belt

An important clue to the history of the early outer solar system is the dynamical excitation of the Kuiper belt. While eccentricities and inclinations of resonant and scattered objects are expected to have been affected by interactions with Neptune, those of the classical objects should have suffered no such excitation. Nonetheless, the confirmed classical belt objects have an inclination range up to at least 32° and an eccentricity range up to 0.2, significantly higher than expected from a primordial disk, even accounting for mutual gravitational stirring.

The observed distributions of eccentricities and inclinations in the Kuiper belt are highly biased. High-eccentricity objects have closer approaches to the Sun and thus become brighter and more easily detected. High-inclination objects spend little time at the low latitudes at which most surveys take place, while low-inclination objects spend zero time at the high latitudes where some searches have occurred. (Latitude and inclination are defined with respect to the

invariable plane, which is a better representation for the plane of the Kuiper belt than is the ecliptic.)

Determination of the eccentricity distribution of the Kuiper belt requires disentanglement of eccentricity and semimajor axis, which is only possible for objects with well-determined orbits for which a well-characterized sample of sufficient size is not yet available. Determination of the inclination distribution, however, is much simpler because the inclination of an object is well determined even after a small number of observations, and the latitude of discovery of each object is a known quantity. Using these facts, *Brown* (2001) developed general methods for debiasing object discoveries to discern the underlying inclination distribution. The simplest method removes the latitude-of-discovery biases by considering only objects discovered within 1° of the invariable plane equator and weights each object by sin(i), where i is the inclination of each object, to account for the proportional fraction of time that objects of different inclination spend at the equator (strictly speaking, one should use only objects found precisely at the equator; expanding to 1° around the equator greatly increases the sample size while biasing the sample slightly against objects with inclinations between 0° and 1°). An important decision to be made in constructing this inclination distribution is the choice of which objects to include in the sample. One option is to use only confirmed classical objects, i.e., those that have been observed at least two oppositions and for which the orbit is reasonably assured of fitting the definition of the classical Kuiper belt as defined above. The possibility exists that these objects are biased in some way against unusual objects that escape recovery at a second opposition because of unexpected orbits, but we expect that this bias is likely to be in the direction of underreporting high-inclination objects. On the other hand, past experience has shown that if we use all confirmed and unconfirmed classical bodies, we pollute the sample with misclassified resonant and scattered objects, which generally have higher inclinations and therefore artificially inflate the inclination distribution of the classical belt. We therefore chose to use only confirmed classical belt bodies, with the caveat that some high-inclination objects might be missing. Figure 3 shows the inclination distribution of the classical Kuiper belt derived from this method. This method has the advantage that it is simple and model independent, but the disadvantage that it makes no use of the information contained in high-latitude surveys where most of the high-inclination objects are discovered. For example, the highest-inclination classical belt body found within 1° of the equator has an inclination of 10.1°, while an object with an inclination of 31.9° has been found at a latitude of 11.2°. The two high-inclination points in Fig. 3 attempt to partially correct this deficiency by using discoveries of objects between 3° and 6° latitude to define the high-inclination end of the inclination distribution, using equation (3) of *Brown* (2001). Observations at these latitudes miss all objects with lower inclinations, but we can linearly scale the high-latitude distribution to match the low-latitude distribution in the region where they are both valid from 6° to 10° and retrieve a correctly relatively calibrated high-inclination distribution.

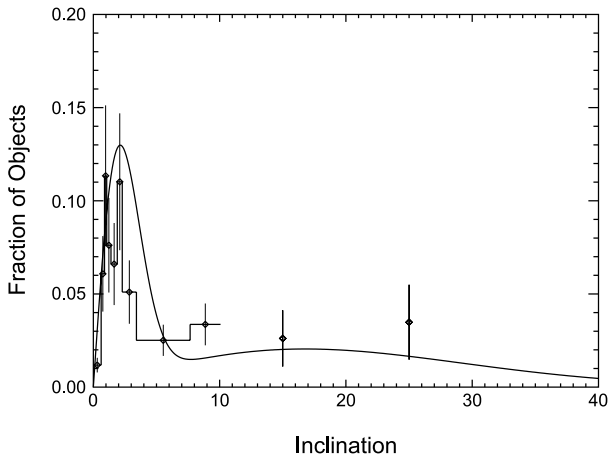

Fig. 3. The inclination distribution of the classical Kuiper belt. The points with error bars show the model-independent estimate constructed from a limited subset of confirmed classical belt bodies, while the smooth line shows the best-fit two-population model.

Brown (2001) developed a more general method to use all objects simultaneously by comparing inclinations of all objects to those found from Monte Carlo observations of simple model inclination distributions at the latitudes of discovery. The simplest reasonable model distribution has a form where f(i)di, the number of objects between inclinations i and i + di, is proportional to $\sin(i)\ \exp(-i^2/2\sigma^2)$ di where σ is a measure of the excitation of the population. The resonant and scattered objects are both well fit by this functional form with $\sigma = 10° \pm 2°$ and $20° \pm 4°$ respectively. The best single Gaussian fit for the confirmed classical belt objects can be ruled out at a high level of confidence; the observed inclination distribution of the classical Kuiper belt is more complex than can be described by the simplest model. Guided by Fig. 3, we make the assumption that the inclination distribution between about 0° and 3° appears adequately described by a single Gaussian times sine inclination, and search for a functional form to describe the higher-inclination objects. The next simplest functional form is one with a second Gaussian added to the distribution: $f(i)di = \sin(i)\ [a_1\ \exp(-i^2/2\sigma_1^2) + a_2\ \exp(-i^2/2\sigma_2^2)]di$. The best fit to the two-Gaussian model, found by modeling the latitudes and inclinations of all confirmed classical belt objects, has parameters $a_1 = 96.4$, $a_2 = 3.6$, $\sigma_1 = 1.8$, and $\sigma_2 = 12$ and is shown in Fig. 3. For this model ~60% of the objects reside in the high-inclination population.

A clear feature of this modeled distribution is the presence of distinct high- and low-inclination populations. While *Brown* (2001) concluded that not enough data existed at the time to determine if the two populations were truly distinct or if the model fit forced an artificial appearance of two populations, the larger amount of data now available, and

shown in the model-independent analysis of Fig. 3, confirms that the distinction between the populations is real. The sharp drop around 4° is independent of any model, while the extended distribution to 30° is demanded by the presence of objects with these inclinations.

3.3. Physical Evidence for Two Populations in the Classical Belt

The existence of two distinct classical Kuiper belt populations, which we will call the hot (i > 4°) and cold (i < 4°) classical populations, could be caused in one of two general ways. Either a subset of an initially dynamically cold population was excited, leading to the creation of the hot classical population, or the populations are truly distinct and formed separately. One manner in which we can attempt to determine which of these scenarios is more likely is to examine the physical properties of the two classical populations. If the objects in the hot and cold populations are physically different, it is less likely that they were initially part of the same population.

The first suggestion of a physical difference between the hot and the cold classical objects came from *Levison and Stern* (2001), who noted that the intrinsically brightest classical belt objects (those with lowest absolute magnitudes) are preferentially found with high inclination. *Trujillo and Brown* (2003) have recently verified this conclusion in a bias-independent manner from a survey for bright objects that covered ~70% of the ecliptic and found many hot classical objects but few cold classical objects.

The second possible physical difference between hot and cold classical Kuiper belt objects is their colors, which relates (in an unknown way) to surface composition. Several possible correlations between orbital parameters and color were suggested by *Tegler and Romanishin* (2000) and further investigated by *Doressoundiram et al.* (2001). The issue was clarified by *Trujillo and Brown* (2002), who quantitatively showed that for the classical belt, inclination, and no other independent orbital parameter, is correlated with color. In essence, the low-inclination classical objects tend to be redder than higher-inclination objects. *Hainaut and Delsanti* (2002) have compiled a list of all published Kuiper belt colors that more than doubles the sample of *Trujillo and Brown* (2002). A plot of color vs. inclination for the classical belt objects in this expanded sample (Fig. 4) confirms the correlation between color and inclination. This expanded sample also conclusively demonstrates that no other independent dynamical correlations occur, although the fact that the low-inclination red classical objects also have low eccentricities, and therefore high perihelia, causes an apparent correlation with perihelion distance as well.

More interestingly, we see that the colors naturally divide into distinct low-inclination and high-inclination populations at precisely the location of the divide between the hot and cold classical objects. These populations differ at a 99.9% confidence level. Interestingly, the cold classical population also differs in color from the Plutinos and the scattered objects at the 99.8% and 99.9% confidence level respectively, while the hot classical population appears identical in color to these other populations. The possibility remains, however, that the colors of the objects, rather than being markers of different populations, are actually caused by the different inclinations. *Stern* (2002), for example, has suggested that the higher average impact velocities of the high-inclination objects will cause large-scale resurfacing by fresh water ice, which could be blue to neutral in color. If this hypothesis were correct, however, we would also expect to see correlations between colors and semimajor axis or eccentricity, which also determine impact velocities. These correlations do not exist. We would also expect to see correlations between color and inclination within the hot and cold populations. Again, these correlations do not exist. Finally, we would expect to see correlations between color and inclination or semimajor axis or eccentricity for all populations, not just the classical belt objects. Once again, no such correlations exist. While collisional resurfacing of bodies may indeed affect colors, there is clearly no causal relationship between average impact velocity and color (*Thébault and Doressoundiram,* 2003). In summary, the significant color and size differences between the hot and cold classical objects implies that these two populations are physically different in addition to being dynamically distinct. A confirmation of the surface composition differences between the hot and cold populations could be made with infrared reflectance spectroscopy, but to date no spectrum of a cold classical Kuiper belt object has been published.

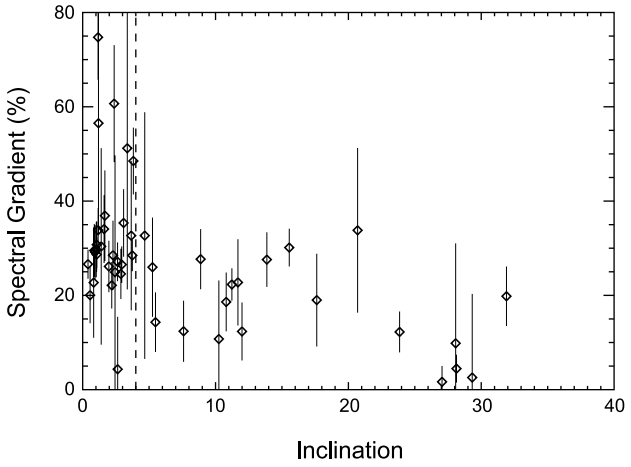

Fig. 4. Color gradient vs. inclination in the classical Kuiper belt. Color gradient is the slope of the spectrum, in % per 100 nm, with 0% being neutral and large numbers being red. The hot and cold classical objects have significantly different distributions of color.

3.4. Radial Extent of the Kuiper Belt

Another important property of interest for understanding the primordial evolution of the Kuiper belt is its radial extent. While initial expectations were that the mass of the

Kuiper belt should smoothly decrease with heliocentric distance — or perhaps even increase in number density by a factor of ~100 back to the level of the extrapolation of the minimum mass solar nebula beyond the region of Neptune's influence (*Stern,* 1996) — the lack of detection of objects beyond about 50 AU soon began to suggest a dropoff in number density (*Dones,* 1997; *Jewitt et al.,* 1998; *Chiang and Brown,* 1999; *Trujillo et al.,* 2001; *Allen et al.,* 2001). It was often argued that this lack of detections was the consequence of a simple observational bias caused by the extreme faintness of objects at greater distances from the Sun (*Gladman et al.,* 1998), but *Allen et al.* (2001, 2002) showed convincingly that for a fixed absolute magnitude, the number of objects with semimajor axis <50 AU was larger than the number >50 AU and thus some density decrease was present.

Determination of the magnitude of the density drop beyond 50 AU was hampered by the small numbers of objects and thus weak statistics in individual surveys. *Trujillo and Brown* (2001) developed a method to use all detected objects to estimate a radial distribution of the Kuiper belt. The method relies on the fact that the heliocentric distance (not semimajor axis) of objects, like the inclination, is well determined in a small number of observations, and that within ~100 AU surveys have no biases against discovering distant objects other than the intrinsic radial distribution and the easily quantifiable brightness decrease with distance. Thus, at a particular distance, a magnitude m_0 will correspond to a particular object size s, but, assuming a power-law differential size distribution, each detection of an object of size s can be converted to an equivalent number n of objects of size s_0 by $n = (s/s_0)^{q-1}$ where q is the differential power-law size index. Thus the observed radial distribution of objects with magnitude m_0, $O(r,m_0)dr$ can be converted to the true radial distribution of objects of size s_0 by

$$R(r,s_0)dr = O(r,m_0)dr\left[\frac{r(r-1)10^{(m-24.55)/5}}{15.60s_0}\right]^{q-1}$$

where albedos of 4% are assumed, but only apply as a scaling factor. Measured values of q for the Kuiper belt have ranged from 3.5 to 4.8 (for a review, see *Trujillo and Brown,* 2001). We will assume the steepest currently proposed value of q = 4.45 (*Gladman et al.,* 2001), which puts the strongest constraints on the existence of distant objects.

Figure 5 shows the total equivalent number of 100-km objects as a function of distance implied by the detection of the known transneptunian objects. One small improvement has been made to the *Trujillo and Brown* (2001) method. The power-law size distribution is only assumed to be valid from 50 to 1000 km in diameter, corresponding to an expected break in the power law at some small diameter (*Kenyon and Luu,* 1999a) and a maximum object's size. The effect of this change is to only use objects between magnitudes 22.7 and 24.8, which makes the analysis only valid from 30 (where a 50-km object would be magnitude 24.8) to 80 AU (where a 1000-km object would be magnitude 22.7). Changes in the maximum object size assumed, s_{max}, are equivalent to changing the outer limit of the validity of the analysis by 80 AU $(s_{max}/1000\text{ km})^{1/2}$. Alternatively, one could further restrict the magnitude limits considered to limit the maximum size while maintaining validity to a particular distance. Different choices of minimum and maximum diameters have little effect on the final result unless extreme values for the maximum are chosen.

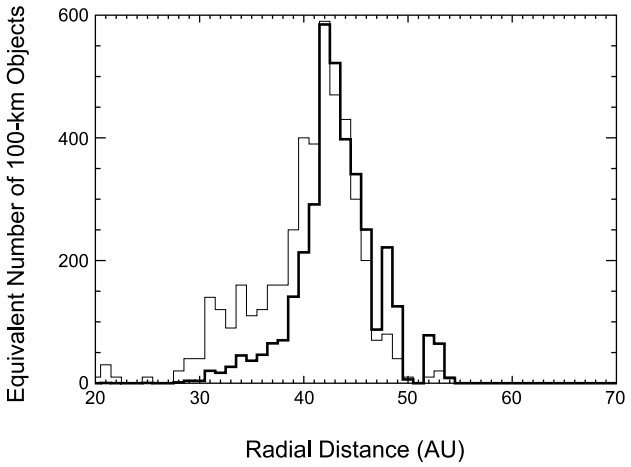

Fig. 5. The radial distribution of the Kuiper belt. The light line shows the observed number of transneptunian objects per AU interval (× 10), while the thick bold line shows the true radial distribution inferred from this observed distribution taking into account biases due to brightness, distance, and size of the object. All discovered transneptunian objects are considered in this analysis, regardless of their dynamical class.

The analysis clearly shows that the known Kuiper belt is a localized increase in number density. Several implicit assumptions go into the above method, but only extreme changes in these assumptions substantially change the results. For example, a change in the object size distribution beyond 43 AU could mimic a drop in object number density, but only if, by 50 AU, the distribution is so extreme that most of the mass is either in a few (undiscovered) large objects or a large number of (too faint) small objects. A physical reason for such a change is not apparent. Likewise, a lowering of albedo beyond 50 AU could make it appear as if there were a drop in number density, but, again, such a lowering is not physically motivated. A change in the inclination distribution beyond 50 AU could have the effect of hiding objects if they are concentrated in low-inclination orbits close to the invariable plane, but repeating the analysis considering only objects found within 1° of the invariable plane still shows the sharp drop. While changing these assumptions could indeed invalidate the analysis method above, the much simpler conclusion is that the number density of the Kuiper belt peaks strongly at 42 AU and quickly drops off beyond that point.

While the *Trujillo and Brown* (2001) method is good at giving an indication of the radial structure of the Kuiper belt where objects have been found, it is less useful for determining upper limits to the detection of objects where none have been found. A simple extension, however, allows us to easily test hypothetical radial distributions against the known observations by looking at observed radial distributions of all objects found at a particular magnitude m_0 independent of any knowledge of how these objects were found. Assume a true radial distribution of objects R(r)dr and again assume the above power law differential size distribution and maximum size. For magnitudes between m and m + dm, we can construct the expected observed radial distribution of all objects found at that magnitude, o(r,m)drdm, by

$$o(r,m)drdm = R(r)dr\left[\frac{r(r-1)10^{(m-24.55)/5}}{15.6s_0}\right]^{-q+1} dm$$

where r ranges from that where the object of brightness m has a size of 50 km to that where the object of brightness m has a size of s_{max}. The overall expected observed radial distribution is then simply the sum of o(r,m) over the values of m corresponding to all detected objects. We can then apply a K-S test to determine the probability that the observed radial distribution could have come from the modeled radial distribution. We first apply this test to determine the magnitude of the dropoff beyond 42 AU. Standard assumptions about the initial solar nebula suggest a surface density drop off of $r^{-3/2}$. Figure 6 shows the observed radial distribution of objects compared to that expected if the surface density of objects dropped off as $r^{-3/2}$ beyond 42 AU. This distribution can be ruled out at the many-sigma levels. Assuming that the surface density drops as some power law, we model a range of different distribution $r^{-\alpha}$ and find a best fit of $\alpha = 11 \pm 4$ where the error bars are 3σ. This radial decay function should presumably hold up to ~60 AU, beyond which we expect to encounter a much flatter distribution due to the scattered disk objects.

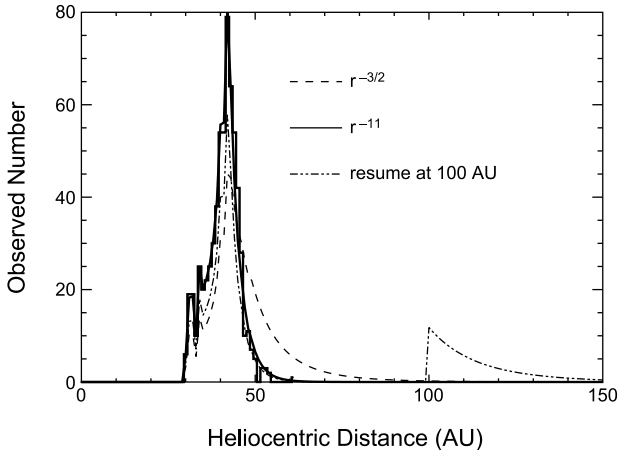

Fig. 6. The observed radial distribution of Kuiper belt objects (solid histogram) compared to observed radial distributions expected for models where the surface density of Kuiper belt objects decreases by $r^{-3/2}$ beyond 42 AU (dashed curve), where the surface density decreases by r^{-11} beyond 42 AU (solid curve), and where the surface density at 100 AU increases by a factor of 100 to the value expected from an extrapolation of the minimum mass solar nebula (dashed-dotted curve).

It has been conjectured that beyond some range of Neptune's influence the number density of Kuiper belt objects could increase back up to the level expected for the minimum mass solar nebula (*Stern,* 1996; see section 3.1). We therefore model a case where the Kuiper belt from 42 to 60 AU falls off as r^{-11} but beyond that the belt reappears at a certain distance δ with a number density found by extrapolating the $r^{-3/2}$ power law from the peak density at 42 AU and multiplying by 100 to compensate for the mass depletion of the classical belt (Fig. 6). Such a model of the radial distribution of the Kuiper belt can be ruled out at the 3σ level for all δ less than 115 AU (around this distance biases due to the slow motions of these objects also become important, so few conclusions can be drawn from the current data about objects beyond this distance). If the model is slightly modified to make the maximum object mass proportional to the surface density at a particular radius, a 100-times resumption of the Kuiper belt can be ruled out inside 94 AU. Similar models can be made where a gap in the Kuiper belt exists at the presently observed location but the belt resumes at some distance with no extra enhancement in number density. These models can be ruled out inside 60 AU at a 99% confidence level.

While all these results are necessarily assumption dependent, several straightforward interpretations are apparent. First, the number density of Kuiper belt objects drops sharply from its peak at around 42 AU. Second, a distant Kuiper belt with a mass approaching that of the minimum mass solar nebula is ruled out inside at least ~100 AU. And finally, a resumption of the Kuiper belt at a density of about 1% expected from a minimum mass solar nebula is ruled out inside ~60 AU.

4. PRIMORDIAL SCULPTING OF THE KUIPER BELT

The previous section makes it clear than the Kuiper belt has lost its accretional disklike primordial structure, sometime during the solar system history. The goal of modelers is to find the scenario, or the combination of compatible scenarios, that can explain how the Kuiper belt acquired the structural properties discussed above. Achieving this goal would probably shed light on the primordial architecture of the planetary system and its evolution.

Several scenarios have been proposed so far. Some of the Kuiper belt properties discussed in section 3 were not yet known when some of these scenarios have been first presented. Therefore in the following — going beyond the original analysis of the authors — we attempt a critical reevaluation of the scenarios, challenging them with all the aspects enumerated in the previous section. We divide the proposed

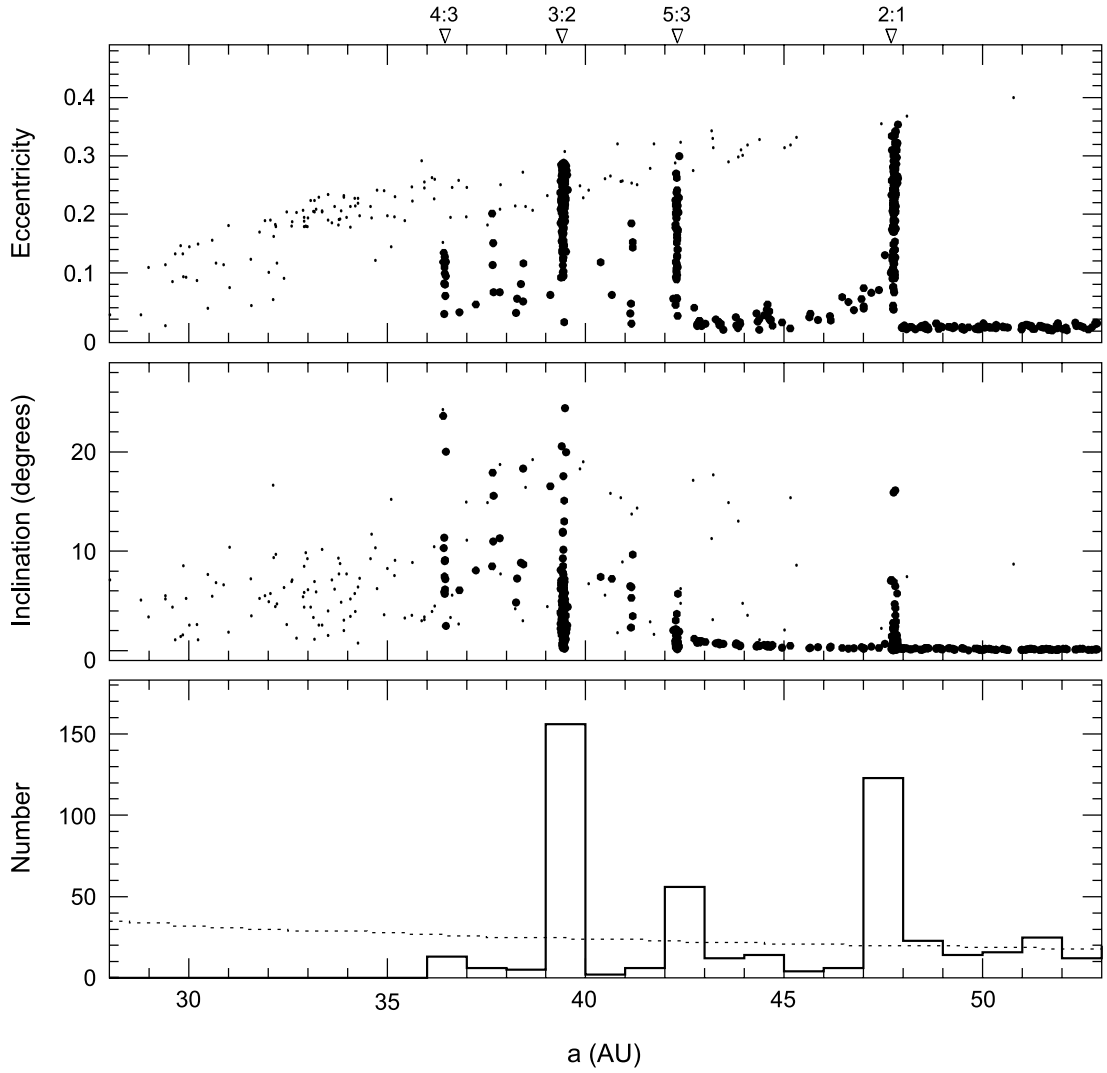

Fig. 7. Final distribution of the Kuiper belt bodies according to the sweeping resonances scenario (courtesy of R. Malhotra). The simulation is done by numerical integrating, over a 200-m.y. timespan, the evolution of 800 test particles on initial quasicircular and coplanar orbits. The planets are forced to migrate (Jupiter: –0.2 AU; Saturn: 0.8 AU; Uranus: 3 AU; Neptune: 7 AU) and reach their current orbits on an exponential timescale of 4 m.y. Large solid dots represent "surviving" particles (i.e., those that have not suffered any planetary close encounters during the integration time); small dots represent the "removed" particles at the time of their close encounter with a planet. In the lowest panel, the solid line is the histogram of semimajor axis of the "surviving" particles; the dotted line is the initial distribution.

scenarios in three groups: (1) those invoking sweeping resonances, which offer a view of gentle evolution of the primordial solar system; (2) those invoking the action of massive scatterers (lost planets or passing stars), which offer an opposite view of violent and chaotic primordial evolution; and (3) those aimed at building the Kuiper belt as the superposition of two populations with distinctive dynamical histories, somehow combining the scenarios in groups (1) and (2).

4.1. Resonance Sweeping Scenarios

Fernández and Ip (1984) showed that, while scattering primordial planetesimals, Neptune should have migrated outward. *Malhotra* (1993, 1995) realized that, following Neptune's migration, the mean-motion resonances with Neptune also migrated outward, sweeping the primordial Kuiper belt until they reached their present position. From adiabatic theory (*Henrard,* 1982), most of the Kuiper belt objects swept by a mean-motion resonance would have been captured into resonance; they would have subsequently followed the resonance in its migration, while increasing their eccentricity. This model accounts for the existence of the large number of Kuiper belt objects in the 2:3 mean-motion resonance with Neptune (and also in other resonances) and explains their large eccentricities (see Fig. 7). Reproducing the observed range of eccentricities of the resonant bodies requires that Neptune migrated by 7 AU.

Malhotra's (1993, 1995) simulations also showed that the bodies captured in the 2:3 resonance can acquire large inclinations, comparable to that of Pluto and other objects. The mechanisms that excite the inclination during the capture process have been investigated in detail by *Gomes* (2000). The author concluded that, although large inclinations can be achieved, the resulting proportion between the number of high-inclination vs. low-inclination bodies and their distribution in the eccentricity vs. inclination plane do not reproduce the observations very well.

The mechanism of adiabatic capture into resonance requires that Neptune's migration happened very smoothly. If Neptune had encountered a significant number of large bodies (1 $M_{\oplus}$ or more), its jerky migration would have jeopardized capture into resonances. *Hahn and Malhotra* (1999), who simulated Neptune's migration using a disk of lunar- to martian-mass planetesimals, did not obtain any permanent capture. The precise constraints set by the capture process on the size distribution of the largest disk's planetesimals have never been quantitatively computed, but they are likely to be severe.

In the mean-motion resonance sweeping model the eccentricities and inclinations of the nonresonant bodies are also excited by the passage of many weak resonances, but the excitation that does occur is too small to account for those observed (compare Fig. 7 with Fig. 1). Some other mechanism (like those discussed below) must also have acted to produce the observed overall orbital excitation of the Kuiper belt. The question of whether this other mechanism acted before or after the resonance sweeping and capture process is unresolved. Had it occurred afterward, it would have probably ejected most of the previously captured objects from the resonances (not necessarily a problem if the number of captured bodies was large enough). Had it happened before, then the mean-motion resonances would have had to capture particles from an excited disk. Another long-debated question concerning the sweeping model is the relative proportion between the number of bodies in the 2:3 and 1:2 resonances. The original simulations by Malhotra indicated that the population in the 1:2 resonance should be comparable to — if not greater than — that in the 2:3 resonance. This prediction seemed to be in conflict with the absence of observed bodies in the former resonance at that time. *Ida et al.* (2000a) showed that the proportion between the two populations is very sensitive to Neptune's migration rate and that the small number of 1:2 resonant bodies, suggested by the lack of observations, would just be indicative of a fast migration (10^5–10^6 yr timescale). Since then, five objects have been discovered in or close to the 1:2 resonance (given orbital uncertainties it is not yet possible to guarantee that all of them are really inside the resonance). There is no general consensus on the debiased ratio between the populations in the 2:3 and 1:2 resonances, because the debiasing is necessarily model-dependent and the current data on the population of the 1:2 resonance are sparse. *Trujillo et al.* (2001) estimated a 2:3 to 1:2 ratio close to 1/2, while *Chiang and Jordan* (2002) obtained a ratio closer to 3. *Chiang and Jordan* (2002) also noted that the positions of the five potential 1:2 resonant objects are unusually located with respect to a reference frame rotating with Neptune, which may also have implications for migration rates and capture mechanisms.

The migration of secular resonances could also have contributed to the excitation of the eccentricities and inclinations of Kuiper belt bodies. Secular resonances occur when the precession rates of the orbits of the bodies are in simple ratio with the precession rates of the orbits of the planets. There are several reasons to think that secular resonances could have been in different locations in the past and migrated to their current location at about 40–42 AU. A gradual mass loss of the belt due to collisional activity, the growth of Neptune's mass, and Neptune's orbital migration would have moved the secular resonance with Neptune's perihelion outward. Levison et al. (personal communication, 1997) found that the Kuiper belt interior to 42 AU would have suffered a strong eccentricity excitation. However, the quantitative simulations show that the orbital distribution of the surviving bodies in the 2:3 resonance would not be similar to the observed one: The eccentricities of most simulated bodies would range between 0.05 and 0.1, while those of the observed Plutinos are between 0.1 and 0.3. Also, in this model there is basically no eccentricity and inclination excitation for the Kuiper belt bodies with $a > 42$ AU, in contrast with what is observed.

The dissipation of the primordial nebula would also have caused the migration of the secular resonances. *Nagasawa and Ida* (2000) showed that the secular resonances involving the precession rates of the perihelion longitudes would have migrated from beyond 50 AU to their current position during the nebula dispersion. This could have caused eccentricity excitation of the Kuiper belt in the 40–50 AU region. In addition, if the midplane of the nebula was not orthogonal to the total angular momentum vector of the planetary system, a secular resonance involving the precession rates of the node longitudes would also have swept the Kuiper belt, causing inclination excitation. The magnitude of the eccentricity and inclination excitation depends on the timescale of the nebula dissipation. A dissipation timescale of $\sim 10^7$ yr is required in order to excite the eccentricities up to 0.2–0.3 and the inclinations up to 20°–30°. But if Neptune was at about 20 AU at the time of the nebula disappearance — as required by the mean-motion resonance sweeping model — the disspation timescale should have been $\sim 10^8$ yr, suspiciously lengthy with respect to what is expected from current theories and observations on the evolution of protoplanetary disks. A major failure of the model is that, because only one nodal secular resonance sweeps the belt, all the Kuiper belt bodies acquire orbits with comparably large inclinations. In other words, the model does not reproduce the observed spread of inclinations, nor their bimodal distribution. No correlation between inclination and size or color can be explained either. The same is not true for the eccentricities, because the belt is swept by several perihelion resonances, which causes a spread in the final

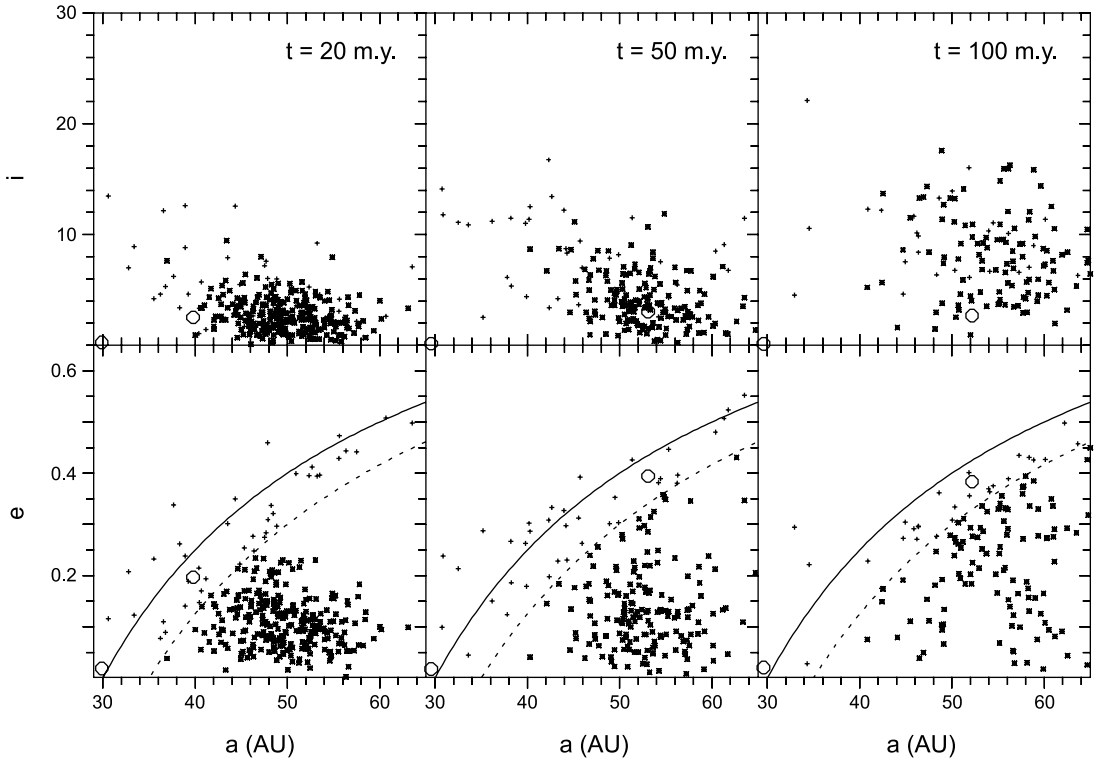

Fig. 8. Snap shots of the Kuiper belt under the scattering action of a 1-$M_\oplus$ planetesimal, itself evolving in the scattered disk. The bold and the dash curves denote q = 30 AU and q = 35 AU respectively. The latter approximately defines the present limit for stability in the Kuiper belt beyond 42 AU, and therefore marks the transition from the classical belt to the scattered disk. The test particles (initially 500, uniformly distributed on circular and coplanar orbits between 35 and 55 AU) are plotted as an asterisk if q > 35 AU, and as a cross otherwise. Neptune and the scattered planetesimal are shown by open circles. From *Petit et al.* (1999).

values. The secular resonance sweeping model cannot explain the existence of significant populations in mean-motion resonances, so that the mean-motion resonance sweeping model would still need to be invoked.

None of the models discussed above can explain the existence of the edge of the belt at ~50 AU.

4.2. Scattering Scenarios

A radically different view has been proposed by *Morbidelli and Valsecchi* (1997), who first proposed that massive Neptune scattered planetesimals (mass on the order of 1 $M_\oplus$), temporarily on Kuiper belt-crossing orbits, could have excited by close encounters the eccentricities and inclinations of the majority of Kuiper belt objects. This idea has been investigated in details by *Petit et al.* (1999), who made direct numerical simulations of the effects of scattered massive planetesimals on test particles representing the initially dynamically cold Kuiper belt. Figure 8 shows snapshots of the status of the Kuiper belt after 20, 50, and 100 m.y. respectively of evolution of an Earth-mass planetesimal in the scattered disk. The test particles (initially 500) were assumed at start on circular and coplanar orbits between 35 and 55 AU. The simulation shows that in the inner belt (a < 40 AU) less than 1% of the bodies are found on what would become "stable orbits" once the massive planetesimal is dynamically removed. The depletion factor in the 40–47 AU region is 74% after 20 m.y., 91% after 50 m.y., and 96% after 100 m.y. This would completely explain the mass deficit of the current belt. However, beyond 50 AU, ~50% of the original test particles are found on "stable orbits" (q > 35 AU) after 100 m.y., which is inconsistent with the observed "edge." In general terms, this model implies a quite steep positive gradient of the number density of bodies vs. semimajor axis, which is not observed in the real population. In particular, the relative population in the 2:3 resonance would be much smaller than that (4% of the classical belt population) claimed from observations by *Trujillo et al.* (2001). In *Petit et al.* (1999) simulations the median eccentricity and inclination of the survivors after 100 m.y. are 0.19 and 8.6° in the 40–47 AU region and 0.27 and 7.4° beyond 47 AU. If the eccentricity distribution correctly reproduces the observations, the inclination distribution is not bimodal, and completely misses objects with inclination larger than 20°, in contradiction with the observations. No correlation between inclination and colors could be explained within the framework of this model.

A variant of the Petit et al. scenario has been invoked by *Brunini and Melita* (2002) to explain the apparent edge of the Kuiper belt at 50 AU. They showed with numerical simulations that a Mars-sized body residing for 1 G.y. on an orbit with a ~ 60 AU and e ~ 0.15–0.2 could have scattered into Neptune-crossing orbits most of the Kuiper belt bodies originally in the 50–70 AU range, leaving this region strongly depleted and dynamically excited. Such a massive body should have been a former Neptune-scattered planetesimal that decoupled from Neptune due to the dynamical friction exerted by the initially massive Kuiper belt. The orbital distribution inside ~50 AU is not severely affected by the massive planetesimal once on its decoupled orbit at a ~ 60 AU (see also *Melita et al.,* 2002). However, a strong dynamical excitation could be obtained during the transfer phase, when the massive planetesimal was transported by Neptune encounters toward a ~ 60 AU, similar to what happens in the *Petit et al.* (1999) simulations. Some of the simulations by *Brunini and Melita* (2002) that include this transfer phase lead to an (a, e) distribution that is perfectly consistent with what is currently observed in the classical belt in terms of mass depletion, eccentricity excitation, and outer edge (see, e.g., their Fig. 10). The corresponding inclination distribution is not explicitely discussed, but it is less excited than in the *Petit et al.* (1999) scenario (M. Melita, personal communication, 2002). Similarly, a correlation between inclination and size or color cannot be reproduced by this mechanism, and a distinctive Plutino population is not formed. Finally, our numerical simulations show that a 1-$M_\oplus$ planet in the Kuiper belt cannot transport bodies up to 200 AU or more by gravitational scattering. Therefore, neither the *Petit et al.* (1999) scenario nor that of *Brunini and Melita* (2002) can explain the origin of the orbit of objects such as 2000 CR_{105}.

A potential problem of the Brunini and Melita scenario is that, once the massive body is decoupled from Neptune, there are no evident dynamical mechanisms that would ensure its later removal from the system. In other words, the massive body should still be present, somewhere in the ~50–70-AU region. A Mars-sized body with 4% albedo at 70 AU would have apparent magnitude brighter than 20, so that, if its inclination is small ($i < 10°$), as expected if the body got trapped in the Kuiper belt by dynamical friction, it is unlikely that it escaped detection in the numerous wide-field ecliptic surveys that have been performed up to now, and in particular in that led by *Trujillo and Brown* (2003).

Another severe problem, for both the *Petit et al.* (1999) and *Brunini and Melita* (2002) scenarios — as well as for any other scenario that attempts to explain the mass depletion of the Kuiper belt by the dynamical ejection of a substantial fraction of Kuiper belt bodies to Neptune-crossing orbit — is that Neptune would have migrated well beyond 30 AU. In *Hahn and Malhotra* (1999) simulations, a 50 $M_\oplus$ disk between 10 and 60 AU drives Neptune to ~30 AU. In this process, Neptune interacts only with the mass in the 10–35-AU disk (about 25 $M_\oplus$), and a massive Kuiper belt remains beyond Neptune. But if the Kuiper belt had been excited to Neptune-crossing orbit, then Neptune would have interacted with the full 50-$M_\oplus$ disk and therefore would have migrated much further. [This fact was not noticed by the simulations of *Petit et al.* (1999) and *Brunini and Melita* (2002), because the former considered a Kuiper belt of massless particles and the latter a Kuiper belt whose total mass was only ~1 $M_\oplus$.] To limit Neptune's migration at 30 AU, the total mass of the disk, including the Kuiper belt, should have been significantly smaller. Our simulations show that even a disk of 15 $M_\oplus$ between 10 and 50 AU, once excited to Neptune-crossing orbit, would drive Neptune too far. Therefore, the scenario of a massive body scattered by Neptune through the Kuiper belt is viable only if the primordial mass of the belt was significantly smaller than usually accepted (accounting only for a few Earth masses).

Motivated by the observation that the eccentricity of the classical belt bodies on average increases with semimajor axis (a fact certainly enhanced by the observational biases, which strongly favor the discovery of bodies with small perihelion distances), *Ida et al.* (2000b) suggested that the structure of the classical belt records the footprint of the close encounter with a passing star. In that paper and in the followup work by *Kobayashi and Ida* (2001), the resulting eccentricities and inclinations were computed as a function of a/D, where a is the original body's semimajor axis and D is the heliocentric distance of the stellar encounter, for various choices of the stellar parameters (inclination, mass, and eccentricity). The eccentricity distribution in the classical belt suggested to the authors a stellar encounter at about ~150 AU. The same parameters, however, do not lead to an inclination excitation comparable to the observed one. The latter would require a stellar passage at ~100 AU or less. From Kobayashi and Ida simulations we argue that a bimodal inclination distribution could be possibly obtained, but a quantitative fit to the debiased distribution discussed in section 3.2 has never been attempted. A stellar encounter at ~100 AU would make most of the classical belt bodies so eccentric to intersect the orbit of Neptune. Therefore, it would explain not only the dynamical excitation of the belt (although a quantitative comparison with the observed distributions has never been done) but also its mass depletion, but would encounter the same problem discussed about concerning Neptune's migration.

Melita et al. (2002) showed that a stellar passage at about 200 AU would be sufficient to explain the edge of the classical belt at 50 AU. An interesting constraint on the time at which such an encounter occurred is set by the existence of the Oort cloud. *Levison et al.* (2003) show that the encounter had to occur much earlier than ~10 m.y. after the formation of Uranus and Neptune, otherwise most of the existing Oort cloud would have been ejected to interstellar space and many of the planetesimals in the scattered disk would have had their perihelion distance lifted beyond Neptune, decoupling from the planet. As a consequence, the extended scattered disk population, with $a > 50$ AU and $40 < q < 50$ AU, would have had a mass comparable or larger than that of the resulting Oort cloud, hardly compatible with

the few detections of extended scattered disk objects performed up to now. An encounter with a star during the first million years from planetary formation is a likely event if the Sun formed in a stellar cluster (*Bate et al.,* 2003). At such an early time, presumably the Kuiper belt objects were not yet fully formed (*Stern,* 1996; *Kenyon and Luu,* 1998). In this case, the edge of the belt would be at a heliocentric distance corresponding to a postencounter eccentricity excitation of ~0.05, a threshold value below which collisional damping is efficient and accretion can recover, and beyond which the objects rapidly grind down to dust (*Kenyon and Bromley,* 2002). The edge-forming stellar encounter could not be responsible for the origin of the peculiar orbit of 2000 CR_{105}. In fact, such a close encounter would also produce a relative overabundance of bodies with perihelion distance similar to that of 2000 CR_{105} but with semimajor axis in the 50–200-AU range. These bodies have never been discovered despite the more favorable observational biases. In order that only bodies with a > 200 AU have their perihelion distance lifted, a second stellar passage at about 800 AU is required (*Morbidelli and Levison,* 2003). Interestingly, from the analysis of the Hipparcos data, *Garcia-Sanchez et al.* (2001) concluded that, with the current stellar environment, the closest encounter with a star during the age of the solar system would be at ~900 AU.

4.3. Scenarios for a Two-Component Kuiper Belt

None of the scenarios discussed above successfully reproduce the existence of a cold and a hot population in the classical belt (see section 3.2–3.3) and the correlation between inclination and sizes and colors. The reason is obvious. All these scenarios start with a unique population (the primordial, dynamically cold Kuiper belt). From a unique population, it is very difficult to produce two populations with distinct orbital properties. Even in the case where it might be possible (as in the stellar encounter scenario), the orbital histories of gray bodies cannot differ statistically from those of the red bodies, because the dynamics do not depend on the physical properties. The correlations between colors and inclination can be explained only by postulating that the hot and cold populations of the current Kuiper belt originally formed in distinctive places in the solar system. The scenario suggested by *Levison and Stern* (2001) is that initially the protoplanetary disk in the Uranus-Neptune region and beyond was uniformly dynamically cold, with physical properties that varied with heliocentric distance. Then, a dynamical violent event cleared the inner region of the disk, dynamically scattering the inner disk objects outward. In the scattering process, large inclinations were acquired. Most of these objects have been dynamically eliminated, or persist as members of the scattered disk. However, a few of these objects somehow were deposited in the main Kuiper belt, becoming the hot population of the classical belt currently observed.

Two dynamical scenarios have been proposed so far to explain how planetesimals in the Uranus-Neptune zone could be permanently trapped in the Kuiper belt. *Thommes et al.* (1999) proposed a radical view of the primordial architecture of our outer solar system in which Uranus and Neptune formed in the Jupiter-Saturn zone. In their simulations, Uranus and Neptune were rapidly scattered outward by Jupiter, where the interaction with the massive disk of planetesimals damped their eccentricities and inclinations by dynamical friction; as a consequence, the planets escaped from the scattering action of Jupiter before ejection on hyperbolic orbit could occur. In about 50% of the cases, the final states resembled the current structure of the outer solar system, with four planets roughly at the correct locations. In this scenario Neptune experienced a high-eccentricity phase lasting for a few million years, during which its aphelion distance was larger than the current one. The planetesimals scattered by Neptune during the dynamical friction process therefore formed a scattered disk that extended well beyond its current perihelion distance boundary. When Neptune's eccentricity decreased down to its present value, the large-q part of the scattered disk became "fossilized," being unable to closely interact with Neptune again. This scenario therefore explains how a population of bodies, originally formed in the inner part of the disk, could be trapped in the classical belt. However, the inclination excitation of this population, although relevant, is smaller than that of the observed hot population. This is probably due to the fact that Neptune's eccentricity is rapidly damped, so that the particles undergo Neptune's scattering action for only a few million years, too short a timescale to acquire large inclinations. For the same reason, the "fossilized" scattered disk does not extend very far in semimajor axis, so that objects like 2000 CR_{105} are not produced in this scenario. Also, the high eccentricity of Neptune would destabilize the bodies in the 2:3 resonance, so that the Plutinos could have been captured only after Neptune's eccentricity damping, during a final quiescent phase of radial migration similar to that in *Malhotra*'s (1993, 1995) scenario. Nevertheless, a Plutino population was never formed in the *Thommes et al.* (1999) simulations, possibly because Neptune's migration was too jerky owing to the encounters with the massive bodies used in the numerical representation of the disk.

Gomes (2003) revisited *Malhotra*'s (1993, 1995) model. Like *Hahn and Malhotra* (1999), he attempted to simulate Neptune's migration, starting from about 15 AU, by the interaction with a massive planetesimal disk extending from beyond Neptune's initial position. But, taking advantage of the improved computer technology, he used 10,000 particles to simulate the disk population, with individual masses roughly equal to twice the mass of Pluto, while Hahn and Malhotra used only 1000 particles with lunar to martian masses. In his simulations, during its migration Neptune scattered the planetesimals and formed a massive scattered disk. Some of the scattered bodies decoupled from the planet by decreasing their eccentricity through the interaction with some secular or mean-motion resonance. If Neptune had not been migrating, as in *Duncan and Levison* (1997) integrations, the decoupled phases would have been

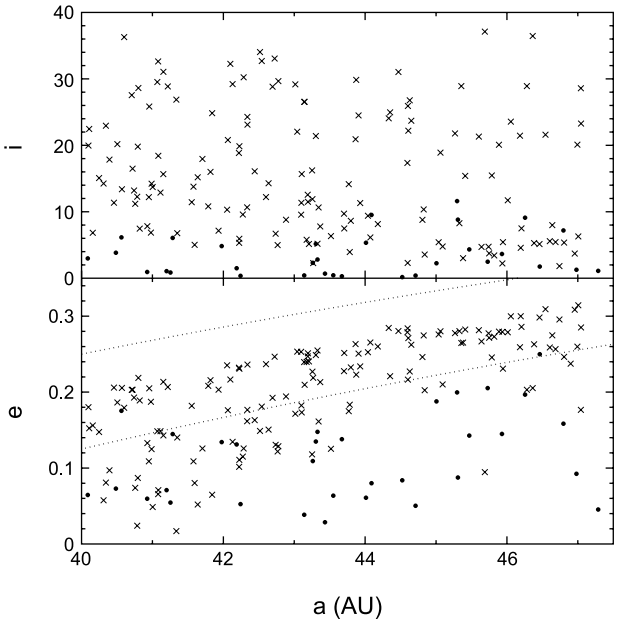

Fig. 9. The orbital distribution in the classical belt according to *Gomes*' (2003) simulations. The dots denote the local population, which is only moderately dynamically excited. The crosses denote the bodies that were originally inside 30 AU. Therefore, the resulting Kuiper belt population is the superposition of a dynamically cold population and of a dynamically hot population, which gives a bimodal inclination distribution comparable to that observed. The dotted curves in the eccentricity vs. semimajor axis plot correspond to q = 30 AU and q = 35 AU. Courtesy of R. Gomes.

transient, because the eccentricity would have eventually increased back to Neptune-crossing values, the dynamics being reversible. But Neptune's migration broke the reversibility, and some of the decoupled bodies managed to escape from the resonances, and remained permanently trapped in the Kuiper belt. As shown in Fig. 9, the current Kuiper belt would therefore be the result of the superposition of these bodies with the local population, originally formed beyond 30 AU and only moderately excited [by the resonance sweeping mechanism, as in *Hahn and Malhotra* (1999)]. Unlike in *Thommes et al.* (1999) simulations, the migration mechanism is sufficiently slow (several 10^7 yr) that the scattered particles have the time to acquire very large inclinations, consistent with the observed hot population. The resulting inclination distribution of the bodies in the classical belt is bimodal, and quantitatively reproduces the debiased inclination distribution computed by *Brown* (2001) from the observations. For the same reason (longer timescale) the extended scattered disk in *Gomes*' (2003) simulations reaches much larger semimajor axes than in *Thommes et al.* (1999) integrations. Although bodies on orbits similar to that of 2000 CR_{105} are not obtained in the nominal simulations, other tests done in *Gomes* (2003) are suggestive that such orbits could be achieved in the framework of the same scenario.

A significant Plutino population is also created in Gomes' simulations. This population is also the result of the superposition of the population coming from Neptune's region with that formed further away and captured by the 2:3 resonance during the sweeping process. Assuming that the bodies' sizes and colors varied in the primordial disk with heliocentric distance, this process would explain why the Plutinos, scattered objects, and hot classical belt objects, which mostly come from regions inside ~30 AU, all appear to have identical color distributions and similar maximum sizes, while only the cold classical population, the only objects actually formed in the transneptunian region, has a different distribution in color and size.

Of all the models discussed in this paper, Gomes' scenario is the one that seems to best account for the observed properties of the classical belt. A few open questions persist, though. The first concerns the mass deficit of the Kuiper belt. In Gomes' simulations about 0.2% of the bodies initially in the Neptune-swept disk remained in the Kuiper belt at the end of Neptune's migration. Assuming that the primordial disk was ~100 $M_\oplus$, this is very compatible with the estimated current mass of the Kuiper belt. But the local population was only moderately excited and not dynamically depleted, so it should have preserved most of its primordial mass. The latter should have been several Earth masses, in order to allow the growth of ~100-km bodies within a reasonable timescale (*Stern,* 1996). How did this local population lose its mass? This problem is also unresolved for the *Thommes et al.* (1999) scenario. The only plausible answer seems to be the collisional erosion scenario, but it has the limitations discussed in section 3.1. Quantitative simulations need to be done. A second problem, also common to the Thommes et al. scenario, is the existence of the Kuiper belt edge at 50 AU. In fact, in neither scenario is significant depletion of the pristine population beyond this threshold obtained. A third problem with *Gomes*' (2003) scenario concerns Neptune's migration. Why did it stop at 30 AU? There is no simple explanation within the model, so Gomes had to artificially impose the end of Neptune's migration by abruptly dropping the mass surface density of the disk at ~30 AU. A possibility is that, by the time that Neptune reached that position, the disk beyond 30 AU had already been severely depleted by collisions. A second possibility is that something (a massive planetary embryo, a stellar encounter, collisional grinding?) opened a gap in the disk at about 30 AU, so that Neptune ran out of material and could not further sustain its migration.

5. CONCLUSIONS AND PERSPECTIVES

Ten years of dedicated surveys have revealed unexpected and intriguing properties of the transneptunian population, such as the existence of a large number of bodies trapped in mean-motion resonances, the overall mass deficit, the large orbital eccentricities and inclinations, and the apparent existence of an outer edge at ~50 AU and a correlation among inclinations, sizes, and colors. Understanding how

the Kuiper belt acquired all these properties would probably constrain several aspects of the formation of the outer planetary system and of its primordial evolution.

Up to now, a portfolio of scenarios have been proposed by theoreticians. None of them can account for all the observations alone, and the solution of the Kuiper belt primordial sculpting problem probably requires a sapient combination of the proposed models. The Malhotra-Gomes scenario on the effects of planetary migration does a quite good job at reproducing the observed orbital distribution inside 50 AU. The apparent edge of the belt at 50 AU might be explained by a very early stellar encounter at 150–200 AU. The origin of the peculiar orbit of 2000 CR_{105} could be due to a later stellar encounter at ~800 AU.

The most mysterious feature that remains unexplained in this combination of scenarios is the mass deficit of the cold classical belt. As discussed in this chapter, the mass depletion cannot be explained by the ejection of most of the pristine bodies to Neptune-crossing orbit, because in this case the planet would have migrated well beyond 30 AU. But the collisional grinding scenario also seems problematic, because it requires a peculiar size distribution in the primordial population and relative encounter velocities that are larger than those that characterize the objects of the cold population; moreover, an intense collisional activity would have hardly preserved the widely separated binaries that are frequently observed in the current population.

A possible solution to this problem has been recently proposed by *Levison and Morbidelli* (2003). In their scenario, the primordial edge of the massive protoplanetary disk was somewhere around 30–35 AU and the entire Kuiper belt population — not only the hot component as in *Gomes*' (2003) scenario — formed within this limit and was transported to its current location during Neptune's migration. The transport process of the cold population was different from the one found by Gomes for the hot population. These bodies were trapped in the 1:2 resonance with Neptune and transported outward within the resonance, until they were progressively released due to the nonsmoothness of the planetary migration. In the standard adiabatic migration scenario (*Malhotra,* 1995) there would be a resulting correlation between the eccentricity and the semimajor axis of the released bodies. However, this correlation is broken by a secular resonance embedded in the 1:2 mean-motion resonance. This secular resonance is generated because the precession rate of Neptune's orbit is modified by the torque exerted by the massive protoplanetary disk that drives the migration. Simulations of this process match the observed (a, e) distribution of the cold population fairly well, while the initially small inclinations are only very moderately perturbed. In this scenario, the small mass of the current Kuiper belt population is simply due to the fact that presumably only a small fraction of the massive disk population was initially trapped in the 1:2 resonance and released on stable nonresonant orbits. The preservation of the binary objects is not a problem because these objects were moved out of the massive disk in which they formed by a gentle dynamical process. The final position of Neptune would simply reflect the primitive truncation of the protoplanetary disk. A bigger problem is the explanation of the different physical properties of the cold and hot populations, because they both originated within 35 AU, although in somewhat different parts of the disk. At the time of this writing, this innovative model has not yet been critically debated within the community of experts. But this scenario offers a simple prediction that will be confirmed or denied by future observations: The edge of the cold classical belt is exactly at the location of the 1:2 resonance.

Kuiper belt science is a rapidly evolving field. New observations change our view of the belt every year. Since the discovery of the first transneptunian object 10 years ago, several review papers have been written, and most of them are already obsolete. No doubt this will also be the fate of this chapter, but it can be hoped that the ideas presented here can continue to guide us in the direction of further understanding of what present observations of the Kuiper belt can tell us about the formation and evolution of the outer solar system.

REFERENCES

Allen R. L., Bernstein G. M., and Malhotra R. (2001) The edge of the solar system. *Astroph. J. Lett., 549,* L241–L244.

Allen R. L., Bernstein G. M., and Malhotra R. (2002) Observational limits on a distant cold Kuiper belt. *Astron. J., 124,* 2949–2954.

Bate M. R., Bonnell I. A., and Bromm V. (2003) The formation of a star cluster: Predicting the properties of stars and brown dwarfs. *Mon. Not. R. Astron. Soc., 339,* 577–599.

Brown M. (2001) The inclination distribution of the Kuiper belt. *Astron. J., 121,* 2804–2814.

Brunini A. and Melita M. (2002) The existence of a planet beyond 50 AU and the orbital distribution of the classical Edgeworth-Kuiper-belt objects. *Icarus, 160,* 32–43.

Chiang E. I. and Brown M. E. (1999) Keck pencil-beam survey for faint Kuiper belt objects. *Astron. J., 118,* 1411–1422.

Chiang E. I. and Jordan A. B. (2002) On the Plutinos and Twotinos of the Kuiper belt. *Astron. J., 124,* 3430–3444.

Chiang E. I., Jordan A. B., Millis R. L., Buie M. W., Wasserman L. H., Elliot J. L., Kern S. D., Trilling D. E., Meech K. J., and Wagner R. M. (2003) Resonance occupation in the Kuiper belt: Case examples of the 5:2 and Trojan resonances. *Astron. J., 126,* 430–443.

Cohen C. J. and Hubbard E. C. (1965) The orbit of Pluto. *The Observatory, 85,* 43–44.

Davis D. R. and Farinella P. (1997) Collisional evolution of Edgeworth-Kuiper belt objects. *Icarus, 125,* 50–60.

Davis D. R. and Farinella P. (1998) Collisional erosion of a massive Edgeworth-Kuiper belt: Constraints on the initial population (abstract). In *Lunar and Planetary Science XIX,* pp. 1437–1438. Lunar and Planetary Institute, Houston.

Dones L. (1997) Origin and evolution of the Kuiper belt. In *From Stardust to Planetesim*als (Y. J. Pendleton and A. G. G. M. Tielens, eds.), p. 347. ASP Conference Series 122.

Doressoundiram A., Barucci M. A., Romon J., and Veillet C. (2001) Multicolor photometry of trans-Neptunian objects. *Icarus, 154,* 277–286.

Duncan M. J. and Levison H. F. (1997) Scattered comet disk and the origin of Jupiter family comets. *Science, 276,* 1670–1672.

Duncan M. J., Levison H. F., and Budd S. M. (1995) The long-term stability of orbits in the Kuiper belt. *Astron. J., 110,* 3073–3083.

Farinella P., Davis D. R., and Stern S. A. (2000) Formation and collisional evolution of the Edgeworth-Kuiper belt. In *Protostars and Planets IV* (V. Mannings et al., eds.), pp. 125–133. Univ. of Arizona, Tucson.

Fernández J. A. and Ip W. H. (1984) Some dynamical aspects of the accretion of Uranus and Neptune — The exchange of orbital angular momentum with planetesimals. *Icarus, 58,* 109–120.

Garcia-Sanchez J., Weissman P. R., Preston R. A., Jones D. L., Lestrade J. F., Latham D. W., Stefanik R. P., and Paredes J. M. (2001) Stellar encounters with the solar system. *Astron. Astrophys., 379,* 634–659.

Gladman B., Kavelaars J. J., Nicholson P. D., Loredo T. J., and Burns J. A. (1998) Pencil-beam surveys for faint trans-Neptunian objects. *Astron. J., 116,* 2042–2054.

Gladman B., Kavelaars J. J., Petit J. M., Morbidelli A., Holman M. J., and Loredo Y. (2001) The structure of the Kuiper belt: Size distribution and radial extent. *Astron. J., 122,* 1051–1066.

Gladman B., Holman M., Grav T., Kaavelars J. J., Nicholson P., Aksnes K., and Petit J. M. (2002) Evidence for an extended scattered disk. *Icarus, 157,* 269–279.

Goldreich P., Lithwick Y., and Sari R. (2002) Formation of Kuiper-belt binaries by dynamical friction and three-body encounters. *Nature, 420,* 643–646.

Gomes R. S. (2000) Planetary migration and Plutino orbital inclinations. *Astron. J., 120,* 2695–2707.

Gomes R. S. (2003) The origin of the Kuiper belt high inclination population. *Icarus, 161,* 404–418.

Hahn J. M. and Malhotra R. (1999) Orbital evolution of planets embedded in a planetesimal disk. *Astron. J., 117,* 3041–3053.

Hainaut O. and Delsanti A. (2002) *MBOSS Magnitude and Colors.* Available on line at http://www.sc.eso.org/~ohainaut/MBOSS/.

Henrard J. (1982) Capture into resonance — An extension of the use of adiabatic invariants. *Cel. Mech., 27,* 3–22.

Ida S., Bryden G., Lin D. N., and Tanaka H. (2000a) Orbital migration of Neptune and orbital distribution of trans-Neptunian objects. *Astrophys. J., 534,* 428–445.

Ida S., Larwood J., and Burkert A. (2000b) Evidence for early stellar encounters in the orbital distribution of Edgeworth-Kuiper belt objects. *Astrophys. J., 528,* 351–356.

Jewitt D. C. and Luu J. X. (1993) Discovery of the candidate Kuiper belt object 1992 QB1. *Nature, 362,* 730–732.

Jewitt D., Luu J., and Chen J. (1996) The Mauna-Kea-Cerro-Totlolo (MKCT) Kuiper belt and Centaur survey. *Astron. J., 112,* 1225–1232.

Jewitt D., Luu J., and Trujillo C. (1998) Large Kuiper belt objects: The Mauna Kea 8K CCD survey. *Astron. J., 115,* 2125–2135.

Kenyon S. J. and Luu J. X. (1998) Accretion in the early Kuiper belt: I. Coagulation and velocity evolution. *Astron. J., 115,* 2136–2160.

Kenyon S. J. and Luu J. X. (1999a) Accretion in the early Kuiper belt: II. Fragmentation. *Astron. J., 118,* 1101–1119.

Kenyon S. J. and Luu J. X. (1999b) Accretion in the early outer solar system. *Astrophys. J., 526,* 465–470.

Kenyon S. J. and Bromley B. C. (2002) Collisional cascades in planetesimal disks. I. Stellar flybys. *Astron. J., 123,* 1757–1775.

Kobayashi H. and Ida S. (2001) The effects of a stellar encounter on a planetesimal disk. *Icarus, 153,* 416–429.

Kuchner M. J., Brown M. E., and Holman M. (2002) Long-term dynamics and the orbital inclinations of the classical Kuiper belt objects. *Astron. J., 124,* 1221–1230.

Kuiper G. P. (1951) On the origin of the solar system. In *Astrophysics* (J. A. Hynek, ed.), p. 357. McGraw-Hill, New York.

Levison H. F. and Duncan M. J. (1997) From the Kuiper belt to Jupiter-family comets: The spatial distribution of ecliptic comets. *Icarus, 127,* 13–32.

Levison H. F. and Morbidelli A. (2003) Pushing out the Kuiper belt. *Nature,* in press.

Levison H. F. and Stern S. A. (2001) On the size dependence of the inclination distribution of the main Kuiper belt. *Astron. J., 121,* 1730–1735.

Levison H. F., Morbidelli A., and Dones L. (2003) Forming the outer edge of the Kuiper belt by a stellar encounter: Constraints from the Oort cloud. *Icarus,* in press.

Lewis J. S. (1995) *Physics and Chemistry of the Solar System.* Academic, San Diego. 556 pp.

Malhotra R. (1993) The origin of Pluto's peculiar orbit. *Nature, 365,* 819–821.

Malhotra R. (1995) The origin of Pluto's orbit: Implications for the solar system beyond Neptune. *Astron. J., 110,* 420–432.

Melita M., Larwood J., Collander-Brown S., Fitzsimmons A., Williams I. P., and Brunini A. (2002) The edge of the Edgeworth-Kuiper belt: Stellar encounter, trans-Plutonian planet or outer limit of the primordial solar nebula? In *Asteroids, Comets and Meteors: ACM 2002 — Proceedings of an International Conference* (B. Warmbein, ed.), pp. 305–308. ESA SP-500, Noordwijk, The Netherlands.

Morbidelli A. and Valsecchi G. B. (1997) Neptune scattered planetesimals could have sculpted the primordial Edgeworth-Kuiper belt. *Icarus, 128,* 464–468.

Morbidelli A. and Levison H. F. (2003) Scenarios for the origin of an extended trans-Neptunian scattered disk. *Icarus,* in press.

Nagasawa M. and Ida S. (2000) Sweeping secular resonances in the Kuiper belt caused by depletion of the solar nebula. *Astron. J., 120,* 3311–3322.

Petit J. M. and Mousis O. (2003) KBO binaries: How numerous were they? *Icarus,* in press.

Petit J. M., Morbidelli A., and Valsecchi G. B. (1999) Large scattered planetesimals and the excitation of the small body belts. *Icarus, 141,* 367–387.

Stern S. A. (1991) On the number of planets in the outer solar system — Evidence of a substantial population of 1000 km bodies. *Icarus, 90,* 271–281.

Stern S. A. (1995) Collisional time scales in the Kuiper disk and their implications. *Astron. J., 110,* 856–868.

Stern S. A. (1996) On the collisional environment, accretion time scales, and architecture of the massive, primordial Kuiper belt. *Astron. J., 112,* 1203–1210.

Stern S. A. (2002) Evidence for a collisonal mechanism affecting Kuiper belt object colors. *Astron. J., 124,* 2297–2299.

Stern S. A. and Colwell J. E. (1997a) Accretion in the Edgeworth-Kuiper belt: Forming 100–1000 km radius bodies at 30 AU and beyond. *Astron. J., 114,* 841–849.

Stern S. A. and Colwell J. E. (1997b) Collisional erosion in the primordial Edgeworth-Kuiper belt and the generation of the 30–50 AU Kuiper gap. *Astrophys. J., 490,* 879–885.

Tegler S. C. and Romanishin W. (2000) Extremely red Kuiper-belt objects in near-circular orbits beyond 40 AU. *Nature, 407,* 979–981.

Thébault P. and Doressoundiram A. (2003) A numerical test of the collisional resurfacing scenario. Could collisional activity explain the spatial distribution of color-index within the Kuiper belt? *Icarus, 162,* 27–37.

Thommes E. W., Duncan M. J., and Levison H. F. (1999) The formation of Uranus and Neptune in the Jupiter-Saturn region of the solar system. *Nature, 402,* 635–638.

Trujillo C. A. and Brown M. E. (2001) The radial distribution of the Kuiper belt. *Astrophys. J., 554,* 95–98.

Trujillo C. A. and Brown M. E. (2002) A correlation between inclination and color in the classical Kuiper belt. *Astrophys. J., 566,* 125–128.

Trujillo C. A. and Brown M. E. (2003) The Caltech Wide Area Sky Survey: Beyond (50000) Quaoar. In *Proceedings of the First Decadal Review of the Edgeworth-Kuiper Belt Meeting in Antofagasta, Chile,* in press.

Trujillo C. A., Jewitt D. C., and Luu J. X. (2001) Properties of the trans-Neptunian belt: Statistics from the Canada-France-Hawaii Telescope Survey. *Astron. J., 122,* 457–473.

Weidenschilling S. (2002) On the origin of binary transneptunian objects. *Icarus, 160,* 212–215.