bered that they integrated for less than 0.2 per cent of the age of the Solar System and their result is only a good indication of Helga's long-term stability rather than proof of it.

Jupiter's gravitational pull is responsible for the creation of a number of large, chaotic regions in the asteroid belt, primarily at orbital radii where the period is a simple fraction of Jupiter's period. Large changes in the orbits can occur at such radii and at 2.5 astronomical units (nearly half the distance of Jupiter's orbit), asteroidal material in chaotic zones can be injected into high eccentricity orbits that cross the Earth's path to become meteorites[7]. Therefore chaos in the asteroid belt is usually associated with an absence of material. However, the orbital periods of 522 Helga and Jupiter are in an approximate 7:12 ratio and a peculiarity of the resulting resonance mechanism acts to protect the asteroid from close approaches to Jupiter[1].

All the above examples of stable chaos in the Solar System are associated with resonance. Pluto is in 2:3 resonance with Neptune as well as a variety of more intricate relationships[3], and the planets of the inner Solar System are involved in a number of long-period or secular resonances[6]. The association of stable chaos with such regions is not too surprising as the boundaries of resonance are known to be sites of chaotic motion (stable and otherwise) and this is observed even in the simplest dynamical systems. Regions where adjacent resonances overlap are also likely places to find chaos.

Undoubtedly there are many more examples of chaotic motion in the Solar System to be found and there will be many more numerical integrations to search for them. This branch of celestial mechanics has almost become an experimental science with the analytical results lagging far behind, but perhaps this is to be expected from a problem which lends itself so easily to numerical solutions. Nevertheless this is an area of research which could have an important impact closer to home; the Earth may not undergo large variations in its orbital elements, but it would be comforting to have similar guarantees for all the minor bodies of the Solar System. □

Carl D. Murray is in the Astronomy Unit, Queen Mary and Westfield College, University of London, Mile End Road, London E1 4NS, UK.

1. Milani, A. & Nobili, A. M. *Nature* **357**, 569–571 (1992).
2. Szebehely, V. *Celest. Mech.* **34**, 49–64 (1984).
3. Sussman, G. J. & Wisdom, J. *Science* **241**, 433 (1989).
4. Wisdom, J. & Holman, M. *Astr. J.* **102**, 1528–1538 (1991).
5. Laskar, J. *Nature* **338**, 237–238 (1989).
6. Laskar, J. *Icarus* **88**, 266–291 (1990).
7. Wisdom, J. *Nature* **315**, 731–733 (1985).

PROTEINS

One thousand families for the molecular biologist

Cyrus Chothia

How many families of proteins are there? By putting together the information to be found in papers published over the past few months we can make an initial estimate, and my calculation suggests that the large majority of proteins come from no more than one thousand families.

Proteins are clustered into families whose members have diverged from a common ancestor and so have similar folds; they also usually have similar sequences and functions. This classification is central to our understanding of how life has evolved, and makes the elucidation and definition of such families one of the principal concerns of molecular biology. With the publication of the first results of the yeast, nematode and human genome projects[1–5], and the availability of two new analyses of known protein structures[6,7], we can begin to put that endeavour on a numerical footing.

The recent reports have described the genes present in the whole of yeast chromosome III and part of chromosome III of the nematode *Caenorhabditis elegans*[1,2]. The genes in part of yeast chromosome IX have been determined by B. Barrell, V. Smith and C. Brown (personal communication). In addition to this genome work, large libraries of human and *C. elegans* expressed genes have been sequenced[3–5]. Table 1 shows the total number of genes identified by this work and the proportion of them that have sequences similar to those determined previously.

The six sets of results give a remarkably consistent view: in each case the authors report that close to one-third of the new sequences belong to families that already have members in the sequence databanks.

The number of protein families represented in the sequence databanks can be calculated approximately from the results reported by Sander and Schneider[6] and by Pascarella and Argos[7]. Many of the protein structures that have been determined by X-ray crystallography and NMR are stored in the Brookhaven databank, and Sander and Schneider[6] determined the proportion of the entries in the EMBL/SwissProt sequence databank that are homologous to the sequences of proteins stored at Brookhaven. This work was originally carried out on 1989 versions of the two databanks[6] but Sander and Schneider have repeated their calculations using the current versions (personal communication). They find that 28% of the entries in the sequence databank have at least a 25% residue identity with one of the entries in the structure bank. This means that at least a quarter of the currently known protein sequences belong to protein families for which there are structures in the Brookhaven databank.

The number of protein families repre-

TABLE 1 New gene sequences that are related to previously determined sequences

Genome projects			
Source	Total number of genes	Genes related to those previously determined	Ref.
Caenorhabditis elegans chromosome III (part)	32	14 (44%)	1
Yeast			
chromosome III	182	52–66 (29–36%)	2
chromosome IX (part)	46	15 (33%)	*
Large libraries of expressed genes			
Source	Total number of clones	Clones related to previously determined protein sequences	
Human brain	~1,400†	406 (~30%)	3
Caenorhabditis elegans			
St Louis–Cambridge	1,517	512 (34%)	4
NIH	585	210 (36%)	5

* B. Barrell, V. Smith and C. Brown, personal communication.
† Adams *et al.*[3] sequenced 2,375 cDNA clones, of which 406 are homologous to previously known sequences. They estimate that not more than half of the remainder contain coding sequences.

sented in the Brookhaven databank can be found from the results of Pascarella and Argos[7], who grouped together the structures that share the same fold (that is, have the same secondary structures in the same orientation with the same chain topology). The 254 different proteins in the April 1991 issue of the Brookhaven structure databank have 83 different folds[7]. Now, proteins with the same fold need not be related: they could have the same fold because of the rules that govern secondary structure packings and chain topologies[8]. But in the large majority of the 83 groups the proteins also have sequence, functional and/or genetic similarities that strongly imply that they are descended from a common ancestor. Consideration of the few instances where this is not the case, and of the many additions made over the past year, shows that the Brookhaven databank contains representatives of 120 different protein families.

TABLE 2 Genome projects

Species	Approximate number of genes	Tentative date of completion
Escherichia coli	4,000	1995–98
Yeast	7,000	2000
Caenorhabditis elegans	15,000	2000
Human	50–100,000	2015

So, to summarize, about a third of the sequences present in genomes are related to entries in the current sequence databank, and about a quarter of the current sequences belong to one of 120 protein families. Assuming that there is no strong bias in the databanks, and that the sequence comparisons give an accurate picture of family membership, this would indicate that there are some 1,500 different protein families. The fairly constant proportions of new sequences that correspond to old sequences (Table 1) suggest that the databanks are not strongly biased. However the sequence comparisons used in this work will underestimate the extent to which proteins are related.

Crystallography has made clear that proteins can evolve; so, though they continue to share the same fold, their residue identities are no greater than that of two randomly selected sequences. For example, actin has a sequence that has been strongly conserved in all eukaryotes and apparently had no relatives. Flaherty *et al.*[9], however, show crystallographically that structural and functional similarities of actin and the ATPase fragment of the heat-shock cognate protein (forms of which are found in both prokaryotes and eukaryotes) are so close there can be little doubt that they are descended from a common ancestor. The two proteins contain 375 and 386 residues, respectively, of which only 39 are identical.

Simple sequence comparisons with reasonable thresholds significantly underestimate the extent to which proteins are related (see also ref. 6). The calculation of the number of protein families described here is rather sensitive to such errors; for example, if we assume that the sequence comparisons find 80% of related proteins, and adjust the calculation accordingly, the number of protein families drops to 1,000. From current crystallographic results it seems that sequence comparisons are not this efficient. Thus a conservative view of the evidence that we have at present would be that the large majority of proteins come from no more than a thousand different families.

Extant proteins have been produced from the original set not just by point mutations, insertions and deletions but also by combinations of genes to give chimaeric proteins. This is particularly true of the very large proteins produced in the recent stages of evolution. Many of these are built of different combinations of protein domains that have been selected from a relatively small repertoire. The evolutionary mechanisms involved in their creation have been described by Patthy[10].

In 1991 some 120 new protein structures were determined; in the previous year the number was 85. More than half of these were unrelated, or only very distantly related, to any previously known structure. Soon there will be 200 new structures each year. In Table 2, I list the current genome sequencing projects and the tentative dates of their completion. If most proteins do come from no more than one thousand families, crystallography, NMR and molecular modelling will produce, at least in outline, structures for most proteins in time for the completion of the genome projects. □

Cyrus Chothia is at the MRC Cambridge Centre for Protein Engineering and the Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

1. Sulston, J. *et al. Nature* **356**, 37–41 (1992).
2. Oliver, S. G. *et al. Nature* **357**, 38–46 (1992).
3. Adams, M. D. *et al. Nature* **355**, 632–634 (1992).
4. Waterston, R. *et al. Nature Genet.* **1**, 114–122 (1992).
5. McCombie, W. R. *et al. Nature Genet.* **1**, 124–130 (1992).
6. Sander, C. & Schneider, R. *Proteins* **9**, 56–68 (1991).
7. Pascarella, S. & Argos, P. *Prot. Eng.* **5**, 121–137 (1992).
8. Chothia, C. & Finkelstein, A. V. *A. Rev. Biochem.* **59**, 1007–1039. (1990).
9. Flaherty, K. M., McKay, D. B., Kabsch, W. & Holmes, K. C. *Proc. Natn. Acad. Sci. U.S.A.* **88**, 5041–5045 (1991).
10. Patthy, L. *Curr. Opin. struct. Biol.* **1**, 351–361 (1991).

DAEDALUS

Pure tones

LAST week Daedalus deduced that a molecule must be instantly evaporated by a photon which exactly supplies its latent heat of evaporation. He calculates that, for water at room temperature, photons of 2.71 micrometres would do the trick. For acetone, the resonant wavelength should be 3.66 micrometres, and for alcohol 3.04 micrometres. So, says Daedalus, shine an infrared laser tuned to 3.04 micrometres on a mixture of these three liquids, and only the alcohol will vaporize.

'Resonant distillation' will be a wonderfully direct and elegant method of chemical separation. Simply by stepping an irradiation laser through a set of frequencies, it will evaporate the components from the most complex mixture one at a time. Daedalus hoped to revolutionize the petrochemical industry with it, until he realized that no infrared laser is powerful enough. But for the bench-scale fractionation of liquid mixtures, resonant distillation should be unrivalled.

A properly tuned laser could quickly flash the traces of chloroform from tap water, or benzene from mineral water. It could neatly vaporize dioxin, caffeine, monosodium glutamate or other additives in foodstuffs. It could even lift the alcohol out of wine, while not disturbing its bouquet or the other subtle flavour-components which alone endear this regrettably intoxicating beverage to its connoisseurs.

More cunning still, resonant distillation is an analytical technique. If the liquid mixture is irradiated by audio-modulated infrared, the selected component will evaporate in a rapid sequence of pulses, giving an audible note whose intensity will reveal the concentration of that component. Tune the modulated laser through a range of wavelengths, and each of the volatiles present will sing out when its resonant wavelength is reached. A musical analyst might even irradiate a mixture with a battery of lasers, each tuned to a specific infrared wavelength and modulated by a particular note, and identify its components from the resulting musical chord. Unwanted components could then be 'sung to exhaustion' — irradiated till their answering note died into silence, showing that they had been completely evaporated.

This neat trick still works even if you don't know what the component is. Every chemical product has to be purified from unknown impurities. Once their frequencies have been identified, they could be sung to exhaustion. Only the desired product would remain, singing the strong single note that guarantees its purity. David Jones