Box 1: Charge, Parity and Time

According to CP symmetry, the laws of physics will appear the same if you simultaneously interchange left and right (for example, reflect in a mirror) and change all particles into their antiparticles. Part of the appeal of CP symmetry is that general principles of quantum mechanics and relativity suggest that CPT — which does these transformations, and also reverses the direction of motion of all particles, or equivalently reverses the arrow of time — is a valid symmetry of nature. Assuming this, CP is then equivalent to T (ref. 15). So in testing CP symmetry, one is testing the proposition that the fundamental laws of nature do not distinguish between past and future. (Of course, this is a very different thing from saying that *the world* does not distinguish between past and future.)

The classic laws of gravity and electromagnetism certainly do possess time-reversal or T symmetry, and the search for CP violation was fruitless for many years. Then Cronin, Fitch and colleagues[16] discovered a small effect in kaon decays that is inconsistent with CP symmetry. Specifically, they found that among long-lived neutral kaons, K_L, which usually decay into three pions, about one in 1,000 particles decay into two pions ($\pi\pi$). If CP were exactly valid, that would never occur. (This is not supposed to be obvious — the argument is subtle, but quite firm.)

It has proved very difficult to improve on Cronin and Fitch. Prior to some very recent experiments[4,5,17], the only firm evidence for CP violation was from kaon decays, and all of it was consistent with the 'superweak' proposal that K_L contained a small admixture of the short-lived kaon K_S, which normally decays into two pions.

The latest measurements refine Cronin and Fitch by comparing the two different possibilities for $K_L \rightarrow \pi\pi$: the pions can either be one positive and one negative, or alternatively both electrically neutral. If the superweak idea is correct, these two possibilities will occur in exactly the same ratio as they do in K_S decays. The actual ratio of ratios is defined to be $1 + 6r$. So if $r \neq 0$ the superweak proposal is wrong, and there's more to CP violation than a tainted K_L. **F. W.**

(or T) violation arises only from effects of the third family explains why its effects are ordinarily so small. We usually study matter made out of just the first two families, so the only CP-violating effects occur through the small, indirect influence of virtual heavy particles from the third family. Ambitious plans are afoot to study particles containing *b* quarks directly. When such particles decay all three families are actively in play, and CP-violating effects are expected to be accessible[10].

What does the Standard Model predict for *r*? This has been the object of some controversy. A major review of the subject[11] quotes $r = 7.7 \times 10^{-4}$, but allows for values from 4.2×10^{-4} to 13.7×10^{-4}. The reason for the spread in this prediction is revealing. Uncertainty arises not so much from any defect in fundamental theory or ignorance of fundamental parameters in the Standard Model, but from weaknesses in our ability to calculate its consequences. For example, there is no simple, reliable way to solve the equations of QCD (quantum chromodynamics) with good accuracy. In making their estimates, theorists resort to rather crude models and drastic approximations, which can come to resemble art more than science.

Into the breach have rushed the number-crunchers, creating a bombshell[6]. By solving the equations of QCD numerically, using ultrafast parallel computers, one obtains, in principle, definitive answers. A group of researchers from Columbia University and Brookhaven National Laboratory now reports that the Standard Model predicts *r* to be negative — actually, $-120\ (\pm\ 70) \times 10^{-4}$ — in dire contradiction with the experimental result. Their result reinforces earlier hints of trouble from less definitive numerical work[12]. Moreover, they suggest that the difference between these calculations and previous predictions lies in their treatment of a particular process described by the 'eye graph' shown in Fig. 1.

Because the source of CP violation in the Standard Model is in principle well understood and in some sense small, experiments in CP violation are especially sensitive to physics outside the Standard Model. Indeed, popular supersymmetry models feature so many potential new sources of CP violation that it is somewhat embarrassing that no deviation from the Standard Model has been observed (until perhaps now). Supersymmetry implies the existence of particles even heavier than the particles in the third family, although there is no evidence for them yet.

As always, extraordinary results merit extraordinary scrutiny. If the Columbia/Brookhaven result holds up, the value of *r* will join neutrino oscillations[13] as the first indications of physics beyond the Standard Model. Already there are serious proposals for how it might be tied up with supersymmetry[14]. ■

Frank Wilczek is at the Institute for Advanced Study, School of Natural Sciences, Olden Lane, Princeton, New Jersey 08540, USA.
e-mail: wilczek@sns.ias.edu

1. Maxwell, J. C. *Theory of Heat* 328 (Longmans, Green, London, 1871).
2. Szilard, L. *Z. Phys.* **53,** 840–856 (1929).
3. Maxwell, J. C. *Scientific Papers* Vol. 2 (ed. Niven, W.) 244 (Dover, New York, 1965).
4. Alavi-Harati, A. *et al. Phys. Rev. Lett.* **83,** 22–27 (1999).
5. Debu, P. http://www.cern.ch/NA48
6. Blum, T. *et al.* http://xxx.lanl.gov/abs/hep-lat/9908025
7. Barr, G. *et al. Phys. Lett. B* **317,** 233–242 (1993).
8. Wolfenstein, L. *Phys. Rev. Lett.* **13,** 562–564 (1964).
9. Kobayashi, M. & Maskawa, K. *Prog. Theor. Phys.* **49,** 652–657 (1973).
10. Ciuchini, M. *et al.* http://xxx.lanl.gov/abs/hep-ph/9704274
11. Bosch, S. *et al.* http://xxx.lanl.gov/abs/hep-ph/9904408
12. Pekurovsky, D. & Kilcup, G. http://xxx.lanl.gov/abs/hep-lat/9903025
13. Wilczek, F. *Nature* **394,** 13–15 (1998).
14. Eyal, G., Masiero, A., Nir, Y. & Silvestrini, L. http://xxx.lanl.gov/abs/hep-ph/9908382
15. Weinberg, S. *The Quantum Theory of Fields I* (Cambridge Univ. Press, 1995).
16. Christenson, J., Cronin, J., Fitch, V. & Turlay, R. *Phys. Rev. Lett.* **13,** 138–140 (1964).
17. Abe, F. *et al. Phys. Rev. Lett.* **81,** 5513–5518 (1998).

Genomics

Functional links between proteins

Andrej Šali

Genome-sequencing projects have accomplished a monumental feat in generating complete lists of the proteins that make up multi-subunit assemblies and signalling pathways in various organisms. These assemblies and pathways must now be mapped, and papers by Marcotte *et al.*[1] and Enright *et al.*[2] (pages 83 and 86 of this issue) take a significant step in this direction. The two groups have developed computational methods that associate proteins through properties other than the similarity between their amino-acid sequences. By comparing phylogenetic (evolutionary) profiles and expression patterns, as well as by analysing domain fusions, the new methods identify proteins that are functionally linked through a metabolic pathway, a signalling pathway or a structural complex (Fig. 1, overleaf). About half of the uncharacterized proteins in yeast — roughly a quarter of all yeast proteins — may be partially annotated by these methods[1]. And, because they do not rely on direct sequence similarity, the methods can assign functions to proteins that lack detectable homologues of known function. They will find many applications in genomics, complementing experiments for the large-scale identification of protein function.

The critical information needed to construct useful models of pathways and assemblies is provided by experiment, most importantly by proteomics and structural biology. Proteomics aims to identify and

characterize complete sets of proteins and protein–protein interactions. It involves large-scale approaches such as the yeast two-hybrid system, mass spectrometry, two-dimensional gel electrophoresis and DNA-microarray hybridization[3]. The size and complexity of the task can be appreciated by assuming between five and 50 functional links per protein, resulting in 30,000 to 300,000 links for a single yeast cell. Although experiments have characterized about 30% of yeast proteins, they are sometimes not rapid, inexpensive or complete enough. So there is a need to assign function using computational methods.

Computational methods have traditionally assigned function by sequence similarity to a characterized protein[4,5]. Such annotation is possible because evolution produced families of homologous proteins that share a common ancestor, and thus have similar sequences, structures and, often, functions. Protein comparisons have allowed some insight into the function of another 30% of yeast proteins[6]. However, functional assignment by homology is limited by two factors. First, it can be applied only to proteins with detectable homologues of known function. Second, it is not always clear what functional properties are shared by the matched proteins, especially for the more distant matches.

The new methods developed by Marcotte *et al.*[1] and Enright *et al.*[2] are not subject to these limitations because they do not depend on sequence similarity between uncharacterized proteins and proteins of known function. Instead, they group proteins that are part of the same pathway or assembly (Fig. 1) and define them as being 'functionally linked'. Marcotte *et al.* have applied three different classification schemes to the proteins in the budding-yeast genome: phylogenetic profiles[7], domain-fusion analysis[8] and correlated messenger RNA expression patterns[9]. The domain-fusion analysis was developed independently by Enright *et al.*, using a new clustering algorithm, and applied to three prokaryotic genomes.

Phylogenetic profiling relies on the correlated evolution of proteins[1,7]. The evolution of two proteins is correlated when they share a phylogenetic profile, which is defined as the pattern of a protein's occurrence over a set of genomes[10]. The phylogenetic profile can be calculated rigorously only when several complete genomes are compared. Two proteins that share a similar phylogenetic profile are expected to be functionally linked. So, clustering of proteins based on their phylogenetic profiles can provide information about the function of an uncharacterized protein that is grouped with one or more functionally defined proteins.

The domain-fusion analysis identifies fusion proteins consisting of two non-homologous component proteins found separately in another genome[1,2]. Such component proteins are expected to interact physically with each other. An interface between two interacting component proteins is more likely to evolve when the proteins are fused in a single chain. A well-known example is fusion of the α and β subunits of tryptophan synthetase from bacteria to fungi. In some respects, the domain-fusion analysis is similar to the use of gene clusters for inferring functional links from gene proximity[11,12].

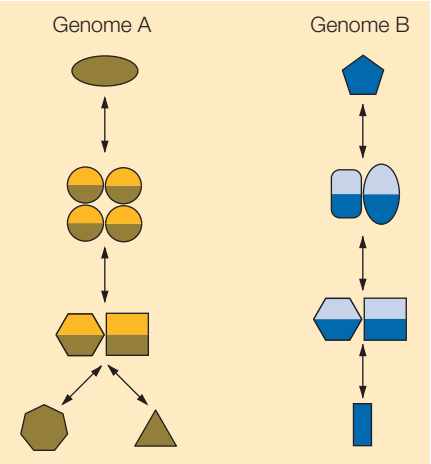

Figure 1 **Functional annotation by computation. Two unrelated pathways consisting of several individual proteins and protein complexes are shown. The sequence and structural similarities of the proteins are indicated by a common shape; the functional links are indicated by a common colour. Some proteins have several links because they are part of the same pathway as well as the same assembly. Traditional annotation methods, which use sequence homology, attempt to annotate by shape. In contrast, the phylogenetic profile[1], domain-fusion analysis[1,2] and expression correlation[1] methods developed by Marcotte *et al.*[1] and Enright *et al.*[2] attempt to annotate by colour.**

Marcotte *et al.* have also grouped yeast proteins by correlating their mRNA expression patterns[1,9]. These patterns were obtained from 97 publicly available DNA chip data sets, which indicate how the expression levels of most yeast proteins change during normal growth, glucose starvation, sporulation and expression of mutant genes. The analysis is based on the expectation that proteins with correlated expression levels over the same series of conditions are functionally linked.

The new functional annotations are often broad, pinning a protein's function down to, say, 'metabolism' or 'transcription'. Even a random pair of proteins have a 50% chance of similar function at such a broad level[1]. But because the annotations are generally derived from a number of linkages, they are three to eight times more informative than random links — comparable, in the best case, to experimental determination of protein–protein interactions[1]. For example, Marcotte *et al.* established new links for MSH6, a DNA-mismatch-repair protein involved in some colorectal cancers, to the PMS1 mismatch-repair family, mutations in which are also tied to human colorectal cancer, the purine-biosynthetic pathway, two RNA-modification enzymes and an uncharacterized protein family, which may now all be investigated in the light of nucleic-acid repair or modification.

How accurate are the annotations, and what percentage of proteins can they cover? These questions can be only partially addressed, because a reference set of functionally linked proteins is not easily available. Marcotte and colleagues assigned a general function to about half of the 2,557 uncharacterized proteins in yeast[13]. They estimated that up to 30% of the pairwise predictions contributing to functional assignments were false, although for the subset of predictions made by two or three of the methods together the rate of false positives decreased to 15%.

Enright *et al.* functionally linked only 215 proteins in three prokaryotic genomes by their domain-fusion analysis, but with very few estimated false positives. Their smaller rate of functional linking seems to be due to the missing links provided by the phylogenetic profiling and mRNA expression methods, which these authors did not use, a stricter definition of fusion events, and the use of fewer proteins to detect fusion. Despite false positives and coarse functional annotation, the computational methods allow experimentalists to concentrate on promising interactions. As more genome data become available, the number of predictions and accuracy of both the domain-fusion analysis and the phylogenetic profiling methods will increase.

The next step is to improve the coverage, accuracy and precision of methods for predicting protein function. This could, in theory, be done by considering three-dimensional structures, because a protein's function is determined more directly by its structure and dynamics than by its sequence. So why have structures not been used as widely as sequences in genomics? There are at least two reasons. First, three-dimensional structures are available for only a fraction of proteins. But this limitation should be reduced by structural genomics within a few years[14]. Structural genomics aims to determine the structures of around 10,000 carefully chosen protein domains, such that all other protein sequences can be modelled with useful accuracy[15]. Second, functional details that can be extracted from structure but not from sequence often depend on the details of that structure in the cellular environment, as well as on its dynamics and energetics, all of which are difficult to obtain by existing experimental and theoretical techniques.

As well as making predictions, bioinformaticians are facing the more practical challenge of making others aware of their predictions. Prediction methods need to be evaluated rigorously and made accessible over the internet[16]. Moreover, varied experimental data and theoretical predictions must be integrated, because no single experimental or computational approach is likely to result in accurate and complete models of protein assemblies and pathways. The latest computational methods for mapping functional links should make a big contribution to such models. ■

Andrej Šali is at the Laboratories of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA.
e-mail: sali@rockefeller.edu

1. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. *Nature* **402,** 83–86 (1999).
2. Enright, A., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. *Nature* **402,** 86–90 (1999).
3. Mendelsohn, A. R. & Brent, R. *Science* **284,** 1948–1950 (1999).
4. Koonin, E. V., Tatusov, R. L. & Galperin, M. Y. *Curr. Opin. Struct. Biol.* **3,** 355–363 (1998).
5. Bork, P. & Koonin, E. V. *Nature Genet.* **18,** 313–318 (1998).
6. Chervitz, S. A. *et al. Nucleic Acids Res.* **27,** 74–78 (1999).
7. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. *Proc. Natl Acad. Sci. USA* **96,** 4285–4288 (1999).
8. Marcotte, E. M. *et al. Science* **285,** 751–753 (1999).
9. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. *Proc. Natl Acad. Sci. USA* **95,** 14863–14868 (1998).
10. Gaasterland, T. & Ragan, M. *J. Microb. Comp. Genomics* **3,** 305–312 (1998).
11. Dandekar, T. *et al. Trends Biochem. Sci.* **23,** 324–328 (1998).
12. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. *Proc. Natl Acad. Sci. USA* **96,** 2896–2901 (1999).
13. Mewes, H. W. *Nucleic Acids Res.* **27,** 44–48 (1999).
14. Burley, S. K. *et al. Nature Genet.* **23,** 151–157 (1999).
15. Sánchez, R. & Šali, A. *Proc. Natl Acad. Sci. USA* **95,** 13597–13602 (1998).
16. Brenner, S. E., Barken, D. & Levitt, M. *Nucleic Acids Res.* **27,** 251–253 (1999).

Geophysics

Latest spin on the core

F. A. Dahlen

Is the Earth's inner core rotating slightly faster than the mantle and crust? This is a question that has perplexed geophysicists since it was first raised five years ago. The latest study comes from Laske and Masters (page 66 of this issue[1]), who have taken a new approach to the problem and conclude that the higher estimates of relative rotation rate can be ruled out.

The centre of the Earth consists of a nearly spherical ball of crystalline iron, 1,220 kilometres in radius; this solid inner core is surrounded by a fluid outer core of molten iron, whose motions are the source of the geomagnetic field. Numerical models of the geodynamo[2,3] suggest that electromagnetic torques might cause the solid inner core to exhibit a slow differential rotation with respect to the silicate mantle and overlying crust, upon which we live. On the other hand, gravitational torques exerted by slight irregularities in the boundary topography and density of the solid inner core and mantle should keep them locked in a state of co-rotation, unless the viscosity of the inner core is low enough to allow it to deform on the requisite timescale[4]. Small but systematic temporal variations in the travel times of seismic waves passing near the centre of the Earth have been interpreted as evidence that the solid inner core is, in fact, rotating faster than the mantle and crust at a rate between $1°\,yr^{-1}$ (ref. 5) and $3°\,yr^{-1}$ (ref. 6). Subsequent seismic travel-time studies have, however, found lower differential rotation rates, ranging from $0.2°–0.3°\,yr^{-1}$ (ref. 7) to essentially zero[8].

Laske and Masters[1] now describe an alternative and preferable seismological procedure for measuring inner-core differential rotation, based upon temporal variations of the splitting of the Earth's large-scale free oscillations, rather than upon the travel times of short-wavelength seismic waves. These latter waves, and the whole-Earth free oscillations (which are analogous to the vibrational modes of a bell), are both generated by large earthquakes. Laske and Masters' analysis, which covers 20 years of terrestrial free-oscillation data, confidently rules out a differential rotation rate as high as $1°\,yr^{-1}$, but is "marginally consistent" with a rate as low as $0.2°–0.3°\,yr^{-1}$.

The principal difficulty in using seismic travel times to probe the inner core is the sensitivity of the measured times to local structural irregularities. It is like trying to measure the rotation rate of a slowly spinning, silvered globe by observing the deflection of a laser beam reflected from its surface. A beam aimed at a perfectly spherical patch will not suffer any deflection no matter how rapidly the globe is spinning; reflection from a large-scale surface irregularity will give rise to a smaller deflection than reflection from a small-scale irregularity. In the seismic problem, the waves travel through the solid inner core rather than reflect from it, and the measured observable is the travel time rather than the angular deflection.

The most persuasive example of temporal variation in seismic travel time is still the one discovered by Song and Richards[5]. Their data came from 30-year records of the arrival times of waves from earthquakes near the South Sandwich Islands at the seismic station in College, Alaska. They found a 0.3-s increase in the time difference between seismic waves that skimmed the inner core and those that passed through it. To convert this $0.01\ s\ yr^{-1}$ temporal variation into their estimated rotation rate, $1°\ yr^{-1}$, Song and Richards assumed that the South Sandwich to College path sampled only a large-scale irregularity of the inner core, namely a slight inclination of its fast axis of anisotropy.

In another study, Su *et al.*[6] used a much larger travel-time data set and two different analysis procedures. But they likewise assumed that only large-scale inner-core asphericity is significant, in deducing a rotation-rate estimate of $3°\,yr^{-1}$. Creager[7] subsequently used a regional seismic network in Alaska to measure the present-day pattern of spatial variation in differential travel time from South Sandwich earthquakes directly.

Box 1: Free oscillations and core rotation

Every normal mode of oscillation of the Earth provides an independent large-scale image of the interior asphericity, known as a splitting function. Loosely speaking, the splitting function $f(\theta,\phi)$ of a mode is a measure of the frequency at which that mode would vibrate if the entire Earth had a spherically symmetrical structure identical to that underlying the point at colatitude θ and longitude ϕ. Only a handful of observably split oscillations 'feel' sufficiently deeply enough into the Earth to be influenced by the inner core. Sharrock and Woodhouse[9] looked for a slow longitudinal shift in the splitting functions of several core-sensitive modes, but their results were equivocal, because of the paucity of suitable earthquake recordings from 20 years ago.

Laske and Masters' approach[1] involved improving the 20-year comparison procedure by first using data from recent earthquakes recorded by the extensive modern global seismic network to determine accurate splitting functions $f = f^{core} + f^{mantle}$ for several core-sensitive modes, and then correcting for the effect of aspherical mantle structure using models of f^{mantle} constrained by independent data. Regarding the mantle as stationary, the authors then sought the rotation that provides the best fit of $f^{rotated\ core} + f^{mantle}$ to the sparser data from past earthquakes. If the inner core is rotating rigidly, then the results from all modes and all past earthquakes should be consistent.

The results achieved with this approach led the authors to conclude that there is either no, or only very slow, differential rotation of the inner core. **F. A. D.**