Library of common protein motifs

SIR—One problem facing molecular biologists is the evaluation of the significance of finding a common amino-acid sequence motif in different proteins. A protein motif[1,2] is a list of allowed residues occuring at specific positions along a sequence and defining a protein's function or structure. These motifs, in which only a few key residues are considered, are increasingly being used to suggest relationships between proteins.

The significance of finding a motif in several proteins is often assessed from the number of motif matches expected by chance in a database search. This number can be estimated from the database frequencies of the residues in a motif (see figure for an example). But as proteins are not random strings of residues, it is not clear that this approach is reliable. Indeed, in conventional homology searches, it is well known that the statistical significance of a sequence relationship is often erroneous[3-5].

Bairoch[2] has established a library, called PROSITE, of 337 protein motifs, based on the 15,409 proteins in the SWISSPROT14 database. Each motif in PROSITE has been defined so it occurs in nearly all members of the protein family in SWISSPROT and generally can be explained biologically — for example, the active-site sequence. In addition, PROSITE details the number of incorrect matches (false positives) when the SWISSPROT14 database is scanned.

Calculated expected number of random matches	Observed number of random matches as given in PROSITE			
	0	1–5	5–10	>10
<0.1	130	1	0	0
0.1–0.5	71	7	1	0
0.5–1.0	18	3	0	0
1.0–5.0	20	14	2	0
5.0–10.0	1	8	2	1
>10	0	0	2	6

PROSITE motifs classified according to the chance expected number of matches and the observed number of matches. Boxed entries, agreement between expected and observed chance matches, representing 73% of the 197 motifs included. Of the 337 motifs in PROSITE, the figure excludes those restricted to a chain terminus; those that yield matches that are classified as unknown; and those that are so general that PROSITE cannot classify true and false matches. For example, one region of DNA polymerases[6] contains the motif (YA)-X-D-T-D-S-(LIVM). In SWISSPROT this motif was correctly identified 24 times and incorrectly found once. There are no matches when it cannot be assessed if the match is a true or false. From the frequencies of residues in SWISSPROT, the probability that one peptide has this motif is: $(0.032+0.077) \times 0.052 \times 0.058 \times 0.052 \times 0.071 \times (0.91+0.054+0.065+0.023) = 2.83 \times 10^{-7}$. Thus the expected number of matches in a scan of the 4.8×10^6 residue SWISSPROT database is 1.4. Data calculated by PROMOT (ref. 7), a program to scan sequence(s) against the PROSITE motif library. PROSITE and SWISSPROT are from the EMBL data library at Heidelberg.

This information provides a benchmark for evaluating motif matches. The number of chance matches given in PROSITE can be compared with the number expected from a calculation of residue frequencies. In the figure the dominance of the boxed entries shows that the expected number of chance matches is a reliable estimate of the observed number of chance matches. This survey excluded motifs that had unknown matches.

From the figure, one obtains a probability of 0.95 (201/210) that a motif match is not a false positive when the expected number (E) of chance matches calcuated from residue frequencies is less than 0.5. Thus if one finds a known motif in a newly determined protein sequence and $E<0.5$, then it is likely that this match detects a biologically meaningful relationship. Clearly, this evaluation is critically dependent on the accuracy in PROSITE of assigning true and false positives to the matches. Some of these assignments are difficult to make but even if there are some errors, the general principle of confidence in the statistics of motif matches should remain.

Motifs must be carefully defined as in PROSITE rather than being developed *a posteriori*. It is not valid to find a weak sequence similarity between two proteins and then arbitrarily to identify common residues to establish a motif that by chance is not expected to occur in a sequence database.

MICHAEL J.E. STERNBERG
Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, PO Box 123, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

1. Hodgman, T.C. *CABIOS* **5**, 1–13 (1989).
2. Bairoch, A. *Prosite: a Dictionary of Protein Sites and Patterns* (Department de Biochimie Medicale, Universite de Geneva, 1990).
3. Dayhoff, M.O., Barker, W.C. & Hunt, L.T. *Meth. Enzym.* **91**, 524–545 (1983).
4. Collins, J.F., Coulson, A. & Lyall, A. *CABIOS* **4**, 67–71 (1988).
5. Sternberg, M.J.E. & Islam, S.A. *Prot. Engng* (in the press).
6. Bernad, A., Zaballos, A., Salas, M. & Blanco, L. *EMBO J.* **6**, 4219–4225 (1987).
7. Sternberg, M.J.E. *CABIOS* (in the press).

HIV and gag

SIR—I have recently encountered a seemingly obvious and, at first glance, simple problem: the relationship of the number of HIV particles to the amount of p24 *gag* gene product in culture medium or in the plasma of seropositive individuals. The measurement of p24 viral core protein by ELISA is a commonly used method to estimate the amount of HIV particles released from infected cells. But no one seems to know precisely how many virions are in 1 ml of a given culture medium containing, for example, 100 pg p24. A survey of the literature on HIV did not yield any clear answer, nor did enquiries to various companies manufacturing p24 ELISA kits. I therefore carried out a series of calculations.

The diameter of mature HIV ranges from 85 to 220 nm with an average size of 100–120 nm (ref. 1): this allows the volume of an individual sphere-shaped virion to be estimated as about 6.97×10^{-16} cm^3 using the formula $4\pi r^3/3$, where r is 55 nm. As the buoyant density of HIV particles[2] is 1.14–1.16 g ml^{-1}, the mass of an individual virion is approximately 8×10^{-16} g.

This, however, is the weight of an entire HIV particle, comprising the weight of core p24. As a rough estimate, the p24 gag protein, the product of one of the nine main genes encoding HIV, must represent at least one-ninth of the total viral body composition[3]. This estimation allocates the share of p24 to 10^{-16} g, which I call p24 constant. To find how many average-sized HIV particles are equivalent to 100 pg of p24, I divide the amount of p24 by the p24 constant. In my example, the number of HIV particles is equal to 100×10^{-12} g/10^{-16} g, or 10^6. It turns out that 100 pg p24 is equivalent to 1 million HIV particles, so 1 million infected lymphocytes in 1 ml medium have to release at least 1 HIV particle per cell to produce up to 100 pg p24 per ml culture supernatant. These numbers agree with an estimation of 10^8 viral particles per ml H9 culture[4] made by counting HIV particles in the electron microscope. H9 cells usually produce about 10 ng p24 which, using my formula, makes exactly 10^8 particles.

Investigations based on the physical mass of HIV and the actual number of infectious particles may be a better alternative to evaluate correctly some properties of HIV. The measurement of infectivity of virus is usually based on the arbitrary units of TCID, which may depend on many variants such as target cell number, susceptibility, viability, doubling time, cellular cycle and so on. Because of this, TCID tends to vary from lab to lab and from experiment to experiment. Contrary to multipartite viruses, one infectious unit of HIV should theoretically correspond to one viral particle. Thus, it may be worthwhile to correlate the infectivity of virus expressed in TCID to the number of real viral particles.

ALDAR S. BOURINBAIAR
The Population Council, Center for Biomedical Research, 1230 York Avenue, New York, New York 10021, USA

1. Lecatsas, G., Taylor, M. B., Lyons, S. F. & Shoub, B. D. *S. Afr. Med. J.* **69**, 793–794 (1986).
2. Levy, J. A. *et al. Science* **225**, 842–845 (1984).
3. Gelderblom, H., Hausmann, E. H. S., Ozel, M., Pauli, G. & Koch, M. A. *Virology* **156**, 171–176 (1987).
4. Popovic, M., Sarngadharan, M. G., Read, E. & Gallo, R. *Science* **224**, 497–500 (1984).