Scientific Correspondence
Nature, Vol. 341, 574-575, October 19, 1989.

Too many leucine zippers?

SIR—The leucine zipper was first proposed[1] as a hypothetical structure facilitating dimer formation of the nuclear proteins C/EBP, Fos, Myc, Jun and GCN4. Its characteristic motif is a periodic repeat of leucines at every seventh position (a heptad repeat) in a segment of 22–29 residues, and its overall composition (in particular, the absence of prolines and glycines, and a high density of ion pairs) is compatible with α-helical secondary structure, suggesting that dimerization is based upon coiled coil interactions[1,2]. However, a distinction has been made between the leucine zipper and the coiled coil interfaces in fibrillar proteins like tropomyosin, keratin and lamins on the basis of the almost exclusive use of leucines in the i+7th positions and the non-conservation of hydrophobic residues in the i+4th positions (in Fos and Myc)[3]. More recently, it has been claimed that many other proteins contain the leucine zipper motif, including the fusion glycoproteins of paramyxoviruses[4], voltage- and ligand-gated ion channels[5] and glucose-transporter glycoproteins[6]. Note, however, that unlike the situation in C/EBP, the motif in the paramyxoviruses entails predominantly charged and uncharged polar residues in the i+4th positions, and the motif in the glucose-transporter proteins entails two substitutions of isoleucine for leucine.

We wish to caution against the too ready acceptance of the significance of leucine heptad repeats found in databank searches. In particular, we address the following questions concerned with the statistical background of such searches. How many leucine heptad repeats might one expect to find merely by virtue of chance? Are leucine heptad repeats relatively more frequent in proteins than heptad repeats of other amino acids (taking into account the biases in residue composition)? Are heptad repeats of leucine more frequent than repeats of other period length?

In the table we list the probability of occurrence of 4 or 5 leucine repeats in random protein sequences of given size and leucine content. Natural protein sequences are not random, but the tabulated probability values may serve as a reference in estimating the statistical significance of the observed repeats. It is seen, for example, that 4 repeats in a 300–500 residue sequence of average leucine content (10%) occur with only 3–4% probability, but that this probability is considerably increased for sequences of higher leucine content or greater length.

We have screened in excess of 450 distinct mammalian proteins of mean length 450 residues and average leucine content 9.7% for occurrences of the motif L-X_6-L-X_6-L-X_6-L, where L is leucine and

Probability of observing a success run of r periodically repeated leucines in a sequence of length n for a given frequency f of leucine in the sequence.

	r = 4			r = 5		
n/f	6.5%	10.0%	13.5%	6.5%	10.0%	13.5%
200	0.003	0.02	0.06	0.001	0.002	0.008
300	0.005	0.03	0.08	0.001	0.003	0.01
500	0.008	0.04	0.13	0.001	0.004	0.02
1000	0.02	0.09	0.25	0.001	0.009	0.04

The probability is closely approximated by $(1 - fx)/[(r + 1 - rx)(1-f)x^{r+1}]$, where x solves $(1 - f)x(1 + fx + .. + f^{r-1}x^{r-1}) = 1$ (ref. 7). A lower bound for the probability of observing an r-repeat of any (not predetermined) amino acid is given by the same formula with values $r - 1$ and f = 5% (0.03–0.06 for r = 4 and n = 300–500).

X is any other amino acid. The motif is found in more than 30 of the sequences, but in only about half of them, including ribophorin I, spectrin and interleukin-3, is X never a proline. Heptad repeats of other amino acids occur much less than half as frequently as the leucine heptads, in agreement with the fact that leucine is by far the most common amino acid over all the proteins analysed. The second most frequent heptad repeat (occurring more than five times as often as expected) contains glutamic acid residues instead of leucine (as, for example, in the i+3th positions of the Fos leucine zipper).

For the same set of proteins, leucine repeats of length 4 with spacing 5 or 7 (rather than 6) are found in about 20 proteins each. Also, spacings 3 and 6 are preferred over spacings 4,5,7 and 8 between (not necessarily nearest) neighbouring leucines. This relative excess of leucine heptad repeats is consistent with the role of leucines in establishing hydrophobic faces of α-helices, including those involved in coiled coil interactions. Interestingly, the same preference for spacing 6 holds for glutamic acid residues, possibly reflecting their role in establishing hydrophilic faces of α-helices.

Thus the leucine repeat by itself occurs widely on both probabilistic and empirical grounds. This abundance undoubtedly reflects the particular structural role of the motif but it also warrants a cautioning note with respect to the prolific citation of leucine zippers, particularly when pattern searches allow for substitutions in the leucine positions. While one might rightly want to accommodate true variations in primary sequence yielding equivalent three-dimensional structures, attention ought to be given to the concomitant increase in the level of false positive occurrences. Based on our probability estimates and data work, we think that perhaps as many as two of every three leucine zipper motifs are simply chance occurrences.

VOLKER BRENDEL
SAMUEL KARLIN
Department of Mathematics,
Stanford University,
Stanford, California 94305, USA

1. Landschulz, W.H., Johnson, P.F. & McKnight, S.L. *Science* **240**, 1759–1764 (1988).
2. O'Shea, E.K., Rutkowski, R. & Kim, P.S. *Science* **243**, 538–542 (1989).
3. Landschulz, W.H., Johnson, P.F. & McKnight, S.L. *Science* **243**, 1681–1688 (1989).
4. Buckland, R. & Wild, F. *Nature* **338**, 547 (1989).
5. McCormack, K *et al. Nature* **340**, 103 (1989).
6. White, M.K. & Weber, M.J. *Nature* **340**, 103–104 (1989).
7. Feller, W. *An Introduction to Probability Theory and its Applications* 3rd edn Vol. 1 325 (Wiley, New York, 1968).