PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

All the motifs in all known proteins

ArticleInfo		
ArticleID	:	3536
ArticleDOI	:	10.1186/gb-2000-1-1-reports004
ArticleCitationID	:	reports004
ArticleSequenceNumber	:	27
ArticleCategory	:	Paper report
ArticleFirstPage	:	1
ArticleLastPage	:	4
ArticleHistory	:	RegistrationDate : 1999–11–2 Received : 1999–11–2 OnlineDate : 2000–3–17
ArticleCopyright	:	BioMed Central Ltd2000
ArticleGrants	:	
ArticleContext	:	130591111

Rachel Brem

Abstract

A

Significance and context

Many evolutionary theories have been developed and tested on satellite DNA, the stretches of non-coding repetitive DNA in eukaryotes. Can repeated sequences in proteins be analyzed in the same way? Marcotte *et al*. try to do this by tabulating all 'protein repeats' from the SWISS-PROT database. The authors define a protein repeat as any piece of protein sequence that appears multiple times with a single protein. This definition then includes both functional motifs like ATP- or calcium-binding sequences, which happen to be used multiple times in a single protein, and also modules like collagen's tripeptide repeats that may exist only for structural reasons. The authors find that proteins in eukaryotes have more repeats than those of prokaryotes, and that the kingdoms have very few repeats in common. They conclude that most repeats arose after the evolutionary split between prokaryotes and eukaryotes. This leads to speculation about why and how eukaryotes developed more repeats.

Key results

Marcotte *et al*. use a previously published repeat-finding algorithm (see Links section) to flag all the proteins in SWISS-PROT that have one or more repeats. From this analysis, they find that 14% of all proteins have repeats and that eukaryotes are three times more likely to have repeats than prokaryotes. Next the authors look at the amino-acid sequences of the repeats themselves. Many are familiar functional or structural motifs, but some are new and of unknown function or fold. Eukaryotes and prokaryotes have almost no repeated sequences in common. This suggests that most repeats evolved after the split between the two kingdoms. Finally, the authors try to work out how repeats arise biochemically, by doing the following analysis. They make a histogram of the lengths of protein repeats - most repeats are short, and very few are long. To this histogram they fit an empirical probability function which contains a parameter E. They interpret E as the energy of making a protein repeat per nucleotide pair. Now they compare their fitted value for E to the true experimental melting energy of DNA per nucleotide pair. Their fitted parameter is two orders of magnitude too small; so Marcotte *et al*. conclude that DNA polymerase slippage, which requires DNA melting, is probably not how protein repeats are formed. From a similar calculation they find that the apparent E is weaker when a protein has

multiple repeats, which may suggest that forming the first repeat is harder than forming additional ones on the same protein.

Links

The authors' Repeat-finding software is available at their website.

Conclusions

A key new conclusion here is that repeats appear more in eukaryotes' proteins than in those from prokaryotes. Marcotte *et al*. speculate on several reasons for this difference. First of all, proteins with repeats may be hard to fold, as they are often composed of many little domains and have no central hydrophobic core. Eukaryotes may handle these repeats better, because these organisms are well equipped to fold difficult proteins. Secondly, the authors argue that repeats may provide an extra source of genomic variability because the DNA replication machinery may make more mistakes on them. Eukaryotes may need repeats more, because they reproduce less often than prokaryotes and so have fewer chances to mutate their genomes.

Reporter's comments

Marcotte *et al*. have a lot of raw data on protein repeats, and they provide only one or two ideas about how to interpret them. This is partly because questions about evolution are hard to quantify. It may also be because the data in this paper are too heterogeneous to be described by many broad theories. Protein repeats are quite different from satellite DNA; all kinds of complicated pressures on a given protein - function, structure and folding - determine whether it has repeating stretches of sequence. When they make their compendium of all protein repeats, Marcotte *et al*. are trying to average over all these individual forces on proteins, but this averaging may not be possible.

Table of links

Journal of Molecular Biology

SWISS-PROT database

Repeat-finding software

References

1. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D: A census of protein repeats. J Mol Biol. 1999, 239: 151-160. 0022-2836