

Searching for Frameshift Evolutionary Relationships Between Protein Sequence Families

Matteo Pellegrini and Todd O. Yeates*

Molecular Biology Institute and UCLA–DOE Laboratory of Structural Biology and Molecular Medicine, University of California, Los Angeles, California

ABSTRACT The protein sequence database was analyzed for evidence that some distinct sequence families might be distantly related in evolution by changes in frame of translation. Sequences were compared using special amino acid substitution matrices for the alternate frames of translation. The statistical significance of alignment scores were computed in the true database and shuffled versions of the database that preserve any potential codon bias. The comparison of results from these two databases provides a very sensitive method for detecting remote relationships. We find a weak but measurable relatedness within the database as a whole, supporting the notion that some proteins may have evolved from others through changes in frame of translation. We also quantify residual homology in the ordinary sense within a database of generally unrelated sequences. *Proteins* 1999;37: 278–283. © 1999 Wiley-Liss, Inc.

Key words: domains; point mutations; statistical significance; evolution; overlapping genes; reading frame

INTRODUCTION

It is well known that the proteins of modern organisms have evolved from a smaller number of ancestral protein sequences. Homologous proteins from different species have diverged through point mutations, insertions, and deletions. Paralogs, which are similar proteins in the same organism, diverged by similar mechanisms after gene duplication events. Evolutionary connections are established among the known proteins by statistical methods of sequence comparison. Typically, dynamic programming algorithms are used to produce an optimal alignment by inserting gaps in the sequences as necessary to bring into register amino acids of the same or similar type.¹ The similarity between amino acid types is quantified in an amino acid substitution matrix that accounts for the probability of point mutations occurring during evolution.^{2,3}

The Swissprot database of protein sequences has been analyzed using these and other techniques, and all homologous protein fragments have been clustered into families.^{4,5} Did the thousands of sequence families found in these databases arise independently, or are there evolutionary relationships between them? One likely possibility is that many of the apparently unrelated families are related

in the usual sense, but have simply diverged beyond recognition by the ordinary mechanisms of point mutations and in-frame insertions and deletions. Here we explore an alternate possibility that some proteins could be related by different evolutionary mechanisms involving changes in frame of translation.

Several lines of reasoning support the possibility that some proteins may have evolved by way of changes in frame of translation. To begin with, the necessary genetic mechanisms are common. Nucleotide insertions and deletions occur frequently during evolution. Although typical analyses of protein sequences reveal only those events that preserve the frame of translation, one might reasonably expect most insertions or deletions to occur out of frame. Other genetic rearrangements could lead to translation from the opposite strand. It might at first appear that such changes in frame of translation during evolution will rarely lead to a new, viable protein that is capable of spontaneously folding. However, recent experiments with random protein sequences⁶ and proteins built from highly reduced amino acid sets⁷ suggest that the criteria for protein folding may be surprisingly permissive.

Furthermore, the degeneracy of the genetic code makes it possible for a single DNA sequence to support simultaneously two overlapping protein sequences in different frames. This is common in viruses, where the situation was first discovered.⁸ There are also special cases in eukaryotic organisms in which alternate RNA splicings give rise to proteins that are in different frames over part of their length (e.g., Grumont et al.⁹). These scenarios further support the notion that frameshifting is a feasible mechanism for protein evolution.

To detect frameshift relationships, we construct five special amino acid substitution matrices that incorporate the combined effects of frameshifts and sequence divergence in the ordinary sense (see Fig. 1 and Materials and Methods). Similar matrices were developed earlier by Claverie.¹⁰ Using these matrices, a set of statistical estimation methods is used to quantify the amount of residual relatedness within a protein sequence database. The present study focuses on well-characterized protein families and asks whether some may be related by changes in frame of translation.

Grant sponsor: Sloan Foundation and Department of Energy post doctoral fellowship.

*Correspondence to: Todd Yeates, Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095-1570.

Received 30 November 1998; Accepted 15 June 1999

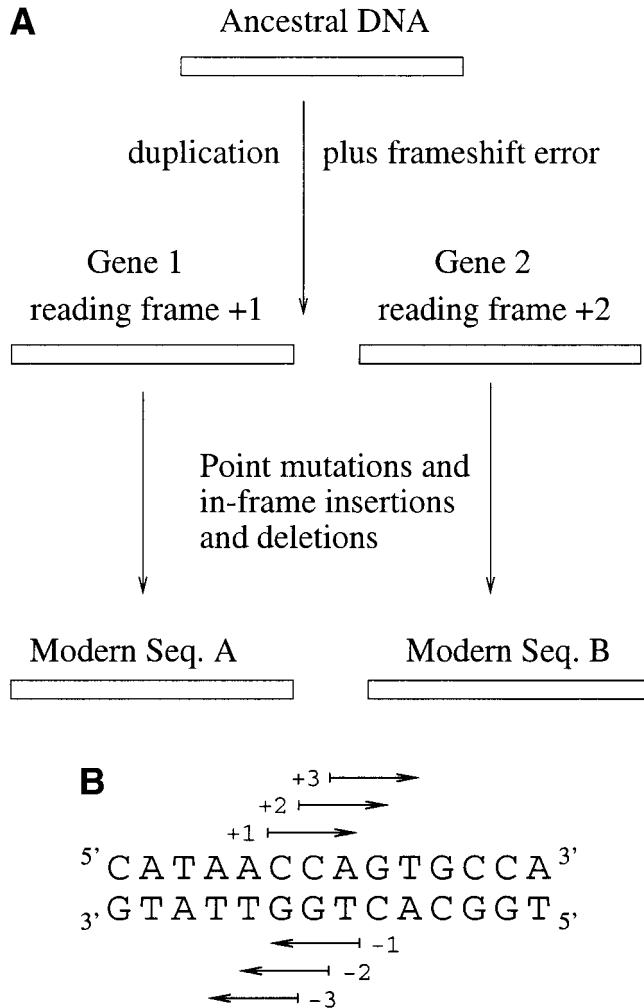


Fig. 1. (A) A schematic representation of evolutionary steps leading to proteins related by a frameshift. An initial duplication of a gene, together with a frameshift error, produces two copies that are translated in different reading frames. The initial event would likely lead to stop codons in the new frame, which would have to be removed through point mutations. The two ancestral proteins then evolve by the usual mechanisms of point mutations and in-frame insertions and deletions, to produce the two modern proteins. (B) Notations for the five alternate frames of translation relative to the standard frame (+1).

MATERIALS AND METHODS

Alignment Algorithm

To find the optimal local alignment between two sequences, we use the Smith–Waterman algorithm.¹¹ This method requires $N \times M$ steps, where N and M are the lengths of the two sequences. Although this method finds the optimal solution to the alignment problem, it is too computationally intensive for an all-versus-all database search.

To render the search more efficient, we use the BLAST¹² approach of aligning only sequences that have a *word* match. A word is defined as a four-amino-acid stretch of the query sequence. Any four-letter segment of the matched sequence that scores above a threshold against a word in the query sequence triggers a nongapped alignment attempt. Sequences that yield statistically significant non-

gapped alignments (see below) are then aligned with a gapped Smith–Waterman alignment. This two-stepped approach allows for efficient searches of databases, with an accuracy that is very close to a full Smith–Waterman search.

Statistical Estimation of Alignment Scores

To detect very distant evolutionary relationships between protein families, it is essential that accurate estimates be obtained for the statistical significance (or P value) of alignment scores between sequences. It is known from preceding work¹³ that the optimal alignment scores between pairs of sequences with fixed amino acid compositions are distributed according to the extreme value distribution:

$$P(S_{n,m} > t) = 1.0 - e^{-\gamma n m p^t} \quad (1)$$

where $S_{n,m}$ is the optimal alignment score between two sequences of lengths n and m , and γ and p are the two parameters that describe the distribution.

The simplest approximation one may use when conducting database searches is to assume that the two parameters γ and p depend only on the total amino acid composition of the database. This expeditious approximation is used in the popular BLAST program,¹² with the added correction that the average length of the optimal local alignment should be subtracted from the values of n and m .

However, the actual values of γ and p may depend rather strongly on the amino acid composition of the two sequences being aligned. As a result, accurate values of these parameters are required to obtain reliable probability estimates. To accurately determine the values of γ and p specific for each pair of sequences, one must calculate the distribution of alignment scores for several hundred random shufflings of the two particular sequences.

We use both methods in progression for assessing statistical significance in our database comparison. The simplest approach, constant γ and p , is used to filter out the lowest scoring pairs. First, the values of γ and p are precomputed once by generating 1,000 pairs of random sequences, with the average amino acid composition of the Prodom database. From these comparisons, the distribution of random alignment scores is computed and the two parameters of the distribution are extracted. The database sequences are then compared by pairwise alignment. Those pairs with a P value less significant than 0.001 are rejected from further consideration.

The sequence pairs that pass this test of significance are examined by the most robust and slowest method: one of the sequences is randomly shuffled 100 times, and an optimal alignment, using the Smith–Waterman algorithm, is computed each time. From this distribution of scores, the γ and p parameters are recomputed and the final statistical significance is assigned.

Construction of Frameshift Matrices

We construct matrices that incorporate the evolutionary mechanism shown in Figure 1. We model the steps of this figure by a three-stage Markov process: (1) a modern protein sequence is related to an ancestral protein by point mutations, insertions/deletions; (2) is subsequently frameshifted; and (3) finally evolves once more.

The point mutation, insertion/deletion component of this process is described by a standard amino acid substitution matrix. To capture this portion of the Markov process, we first convert the BLOSUM62 amino acid substitution matrix³ from a log odds matrix to a conditional probability (or transition) matrix:

$$P_B(i \rightarrow j) = p(j)2^{\text{BLOSUM62}_{ij}/2}. \quad (2)$$

$P(i \rightarrow j)$ is the probability that amino acid i will be replaced by amino acid j through point mutations according to the BLOSUM62 scores. The p_j 's are the abundances of amino acid j and are computed by solving the 20 linear equations given by the normalization conditions that

$$\sum_j P_B(i \rightarrow j) = 1. \quad (3)$$

We then compute the n th power (for nonintegral values of n) of this matrix by calculating the Taylor series expansion of

$$P_B^n(i \rightarrow j) = e^{n \ln(P_B)}. \quad (4)$$

Raising the probability matrix to a power simulates different time lengths of evolution. A power greater than unity produces a matrix that has a more uniform distribution of values, and best relates sequences that have diverged further. An exponent smaller than one has the opposite effect.

The effect of the frameshift is captured in a separate matrix. The frameshift conditional probability matrix, P_{fs} , accounts for the conversion of a particular codon for amino acid i into amino acid j in a different reading frame. For instance, the codon *ACG* codes for threonine. When the complementary strand, *TGC*, is read backwards in the -2 reading frame, *GTX* (where X is the unknown nucleotide of the neighboring codon) codes for valine. Therefore, the -2 frameshift conditional probability matrix would reflect that threonine has a high probability of translating to a valine. In all cases, we use the convention defined in Figure 1B, that the default translation of a nucleotide sequence is the $+1$ frame, the one and two nucleotide insertions are $+2$ and $+3$, and the translations of the complementary strand are -1 , -2 , and -3 .

To compute the five frameshift conditional probability matrices, we tabulate for each reading frame the pairs of amino acids corresponding to the 64 codons. The values of the matrix are then given by

$$P_{fs}(i \rightarrow j) = \frac{p_{fs}(i, j)}{p_{fs}(i)}, \quad (5)$$

where $p_{fs}(i, j)$ is the probability of finding the (i, j) amino acid pair and $p_{fs}(i)$ is the probability of finding the amino acid i . No attempt is made to model the possible effects of codon bias at this stage.

To obtain the full frameshift substitution matrix, we multiply the frameshift conditional probability matrix on the left and right by the n th power of the BLOSUM62 conditional probability matrix,

$$P_{tot}(i \rightarrow j) = \sum_{k, l} P_B^n(i \rightarrow k) P_{fs}(k \rightarrow l) P_B^n(l \rightarrow j). \quad (6)$$

The product of the three matrices accounts for the fact that after a frameshift error occurs, both resulting sequences evolve in the ordinary sense, as embodied in the BLOSUM62 matrix. This matrix is then converted to a log-odds matrix, when used to evaluate scores of sequence alignments:

$$S_{log-odds}(i, j) = \log_2 \frac{P_{tot}(i \rightarrow j)}{p(j)}. \quad (7)$$

In the present study, we choose the value of n for each frame so as to fix the information content in all five matrices to the value of the BLOSUM62 matrix. Information values are computed as

$$I = - \sum_{i, j} p(i, j) S_{log-odds}(i, j), \quad (8)$$

where $p(i, j)$ is the joint probability of amino acid i and j , and $s(i, j)$ is the corresponding log-odds score from the matrix.

The necessary values for n are: 0.35 for -1 , 0.45 for -2 , 0.2 for -3 , and 0.3 for the $+2$ and $+3$ reading frames. This implies that the translation of a codon in -2 reading frame contains the most information. This is so because the wobble positions of the $+1$ and -2 reading frames overlap. Conversely, the -3 reading frame contains the least information. A priori, we would expect the frame of comparison with the greatest information content to have the highest probability of finding homologous pairs of sequences related by frameshift.

The gap penalties associated with each frameshift matrix were chosen to optimize the alignment results in a synthetic database. We generated 100 pairs of sequences. One sequence in the pair resulted from the translation of a random 1,000-nucleotide sequence. The nucleotide sequence was then randomly mutated and gaps were inserted (one gap per 100 mutations) and then translated in a different reading frame to generate the second amino acid sequence of the pair. We then aligned each sequence against each other sequence in the pair database using the appropriate frameshift substitution matrix. The gap penalties were chosen to maximize the differences between P values for true and false positives. The frameshift substitution matrices are available on the Internet at <http://www.doe-mbi.ucla.edu/people/yeates/frameshift>.

RESULTS AND DISCUSSION

Among the possible databases one may choose to analyze, we opted for the Prodom database,⁴ version 34.2, because it contains consensus sequences for each protein domain family. From this database, we extracted 8,823 consensus sequences with an average length of 90 and a maximum length of 399. Only families that contained at least four representative domains were retained. This criterion was imposed in an effort to eliminate missequenced domains from our list. It is known that sequencing errors are quite common,¹⁰ and their presence among our list would lead to spurious frameshift relationships. By selecting only families with four or more representatives, we effectively eliminate this possibility, because it is unlikely that all the proteins in the family have been sequenced incorrectly.

For each alternate reading frame, the database sequences are compared pairwise with the two-step procedure described in Materials and Methods. Each sequence alignment yields a single probability score, so an all-against-all database comparison produces a distribution of roughly 100 million P values. We ask whether the observed distribution of P values differs significantly from the distribution expected by random chance.

To answer this key question, we repeat the same calculation with a shuffled version of our sequence database. Because hexanucleotide biases are known to exist within coding regions, we shuffle the database sequences in blocks of three amino acids to preserve any such bias. We also shuffle sequences in single amino acid blocks. The distribution of P values from all-versus-all alignments in the shuffled database may be treated as a background signal, allowing us to verify whether the true database yields a distribution that is above the noise level. Because of the large number of comparisons, the method of analysis is very sensitive to faint sequence relationships on a database-wide scale.

In a search with the plus-two frameshift matrix, the distribution of P values as a whole is slightly higher for the real versus the shuffled database. When the results are examined in Figure 2, we find an excess of approximately 50 scores with P values less than 10^{-4} . This result suggests that a subset of the modern protein domain families may be distantly related to each other by frameshifts. However, the detectable residual homology between sequence families is at a level that does not permit us to conclude that specific families are unambiguously related, except in a few special cases. In essence, we can accurately measure the degree of residual homology within the database as a whole, although the confidence in each particular relationship may be relatively low.

When we examined the P value distribution in detail, we also found 13 hits with P values well beyond background, as seen in the inset of Figure 2. We analyzed these results in detail and found that all of these clearly matching pairs were viral proteins. When we examined the genomes from which these proteins were translated, we found that they correspond to overlapping viral genes.

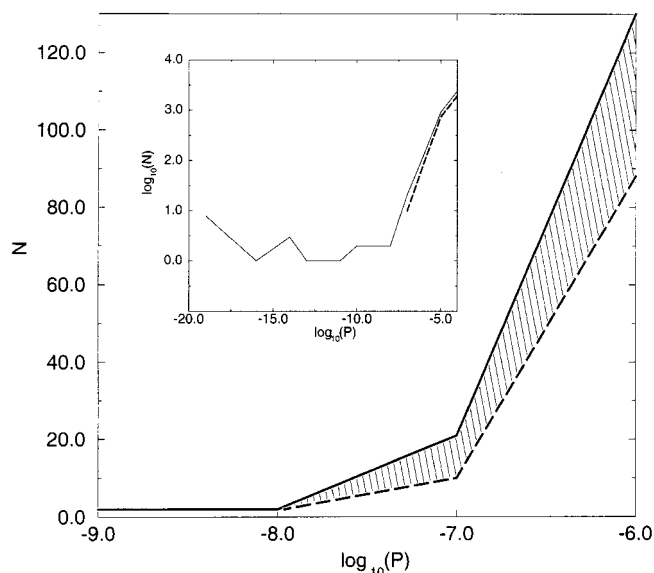


Fig. 2. Distribution of alignment scores in searches for translation from frameshifted reading frames on the coding strand. A plot of the number of sequence pairs versus their P value computed using the +2/+3 frameshift matrix. The solid line is for pairwise comparison of the Prodom database. The dashed line is for pairwise comparisons of the shuffled version of the database. A complete view of the distribution is displayed on a \log - \log plot in the inset. The shaded region represents the number of protein sequences believed to be related by frameshifts.

TABLE I. Number of Frameshifted Sequence Comparisons With $P < 10^{-3}$

Reading frame	Unshuffled	Shuffled ^a	Excess hits
+1	34,689	23,726	10,963
-1	49,128	47,107 (46,862) \pm 45 ^b	2,021
-2	37,795	36,600 (36,673)	1,122
+2	48,487	46,725 (46,299)	1,762
+3	37,637	36,059 (36,362)	1,275

^aValues in parentheses are for sequences shuffled in blocks of three amino acids.

^bThe variance was computed from the results of three separate shufflings of the database. The uncertainty may also be estimated from Poisson counting statistics to be approximately 200. Similar estimates should apply to the other frames of translation.

We repeated the calculation of the P value distribution with the -1, -2, and +3 frameshift matrices. The results of these calculations are summarized in Table I. We find that as in the previous case, the signal from the real database is slightly higher than that from the shuffled database. We show that the total number of sequences with a P value less than 10^{-3} is higher for the real versus the shuffled databases.

The method we have presented may also be used to quantify the degree of residual homology in the ordinary sense within a database. To demonstrate this, we conducted alignments between all the sequences in our filtered Prodom list using the BLOSUM62 matrix. It was recognized by the authors of the Prodom database that residual homology exists between the domain families as a

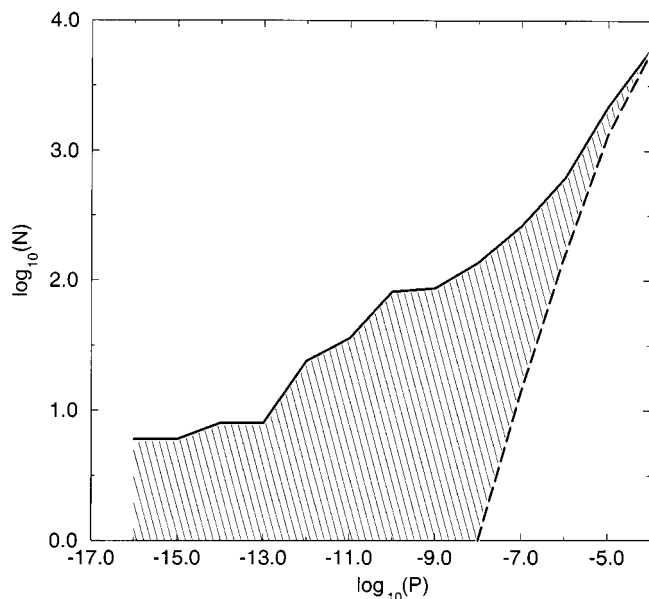


Fig. 3. Residual homology within the Prodom database of distinct sequence families. A \log - \log plot of the number of sequence pairs versus their P value computed using the BLOSUM62 matrix. The thick line is for a pairwise comparison of the Prodom database. The thin line is for a shuffled version of the database, which is linear as expected and gives the distribution of scores expected for a set of unrelated sequences. The shaded region represents the number of protein families believed to be related to each other by distant homology.

consequence of applying too stringent a P value cutoff when using the BLAST program.⁴ This was necessary to avoid generating spurious alignments. By implementing our method we were able to quantify this statement. The results are shown in Figure 3, which depicts the histogram of P values for the true and shuffled databases. A comparison of these distributions demonstrates that, as expected, the signal from the true database is significantly stronger than that from the shuffled database. By subtracting the number of low probability alignments expected for random sequences from the number found for the true sequences, we find that approximately 2,000 sequence pairs yield P values above background. Although some of these sequence pairs contain the same single sequence, a liberal estimate is that 20% of the sequences in this particular database have relatives within the database.

CONCLUSIONS

We have conducted a careful statistical analysis of pairwise sequence comparisons of domains extracted from the Prodom database using five amino acid substitution matrices designed to capture the effects of frameshifts during evolution. Because the sought-after relationships are remote, obvious relationships between specific protein families are generally not seen or anticipated. Instead, we have compared the complete distribution of pairwise probability scores to corresponding distributions obtained with a database of shuffled sequences. The comparison of these

distributions reveals faint relationships between sequences on a database-wide scale.

We find that although the P value distributions are generally similar for the real and shuffled databases, there appears to be a real abundance of low probability scores, particularly in the same-strand frameshift comparisons. The results suggest a probable frameshifted evolutionary relationship between several hundred domain families. Unfortunately, because of the weakness of the signal, it is generally not possible to say with certainty which families might actually be related. A few sequence pairs do produce very significant frameshift scores far above background, but these can all be traced to viral genomes, where the genes coding for the proteins are physically overlapping, primarily on the same strand.

The results presented here are statistical in nature, and there may be other possible explanations for the data. DNA sequencing errors are a cause for concern, but we have circumvented that problem by examining only well-characterized sequence families. The appearance of frameshift similarity might also arise from some peculiarity of the genetic code, and we have attempted to account for this by shuffling the database sequences in blocks, preserving potential hexanucleotide biases. We believe that the simplest explanation is that the signal we observe, even though weak, is caused by true frameshift homology between domain families.

We have also quantified residual homology, in the ordinary sense, between Prodom domain families using the BLOSUM62 matrix. We estimate that approximately 20% of the domain families in this particular database have common ancestors. The statistical approach taken here is generally useful for determining whether further similarities exist in a database and in improving estimates of the number of unique protein families.

The present study focused on the relatively common protein sequence families, which must be relatively old, predating the speciation of the organisms in which they appear. An investigation of newly emerged or ORFan¹⁴ sequences (having no homologues in other species) might provide more evidence for frameshifts during evolution.

ACKNOWLEDGMENTS

We thank Drs. Danny Rice, Robert Grothe, and David Eisenberg for many useful discussions. Part of this work was supported by a Sloan foundation and the Department of Energy postdoctoral fellowship (M.P.).

REFERENCES

1. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J Mol Biol* 1970;48:443-453.
2. Dayhoff MO, Barker WC, Hunt LT. Establishing homologies in protein sequences. *Methods Enzymol* 1983;91:524-545.
3. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;89:10915-10919.
4. Sonnhammer ELL, Kahn D. Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci* 1994;3:482-492.
5. Sonnhammer ELL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 1997;28:405-420.

6. Davidson AR, Lumb KJ, Sauer RT. Cooperatively folded proteins in random sequence libraries. *Nature Struct Biol* 1995;2:856–863.
7. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH. Protein design by binary patterning of polar and nonpolar amino acids. *Science* 1993;262:1680–1685.
8. Barrell BG, Air JM, Hutchinson CA. Overlapping genes in bacteriophage ϕ X174. *Nature* 1976;264:34–36.
9. Grumont RJ, Fecondo J, Gerondakis S. Alternate RNA splicing of murine *nfb1* generates a nuclear isoform of the p50 precursor NF-kappa B1 that can function as a transactivator of NF-kappa B-regulated transcription. *Mol Cell Biol* 1994;14:8460–8470.
10. Claverie J. Detecting frame shifts by amino acid sequence comparison. *J Mol Biol* 1993;234:1140–1157.
11. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;147:195–197.
12. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
13. Waterman MS, Vingron M. Sequence comparison significance and Poisson approximation. *Stat Sci* 1994;9:367–381.
14. Fischer D, Eisenberg D. Finding families for genomic ORFans. *Bioinformatics* 1999;15: in press.