Fundamentals of database searching

Aligning novel sequences with previously characterized genes or proteins provides important insights into their common attributes and evolutionary origins. The principles underlying the computational tools that can be used to evaluate sequence alignments are discussed.

Stephen F. Altschul
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.
altschul@ncbi.nlm.nih.gov

Efficient DNA sequencing methods make it much easier to obtain information on the amino acid sequence of proteins than on their structures or functions. The sequences of homologous proteins can diverge greatly over time, even though the structure or function of the proteins change little. Thus, much can be inferred about an uncharacterized protein when significant sequence similarity is detected with a well-studied protein. This has been a key motivation for the comparison of DNA and protein sequences. Other goals of sequence comparison include phylogenetic reconstruction and the detection of genes and regulatory regions (see the article by David Haussler on pp. 12–15).

Global and local sequence alignment

Alignments provide a powerful way to compare related sequences, but can be used in an attempt to capture different facts. The alignment of two residues could reflect a common evolutionary origin, or could try to represent common structural roles, which might not always be congruent with evolutionary history. Here, I examine the evolutionary view.

Alignments are generally restricted to describing the most common mutations: insertions, deletions and single-residue substitutions. Insertions or deletions are represented by **null characters**, added to one sequence and aligned with letters in the other; substitutions are represented by the alignment of two different letters. Sequences can be compared by either **global** or **local alignment**, depending on the purpose of the comparison (see Fig. 1). Global alignment forces complete alignment of the input sequences, whereas local alignment aligns only their most similar segments. The method used depends upon whether the sequences are presumed to be related over their entire lengths or to share only isolated regions of homology. Although global and local alignment **algorithms** are reasonably similar, the statistics needed to assess their output are very different.

Alignment scores

To select from the vast number of possible alignments, the standard procedure is to assign them scores; the most frequently used convention is that the higher the score the better the alignment. There are many possible definitions of alignment score, but the most common is simply the sum of scores specified for the aligned pairs of letters, and letters with nulls, of which an alignment consists. A **substitution score** is chosen for each pair of letters that can be aligned; the complete set of these scores is called a **substitution matrix** [PAM (Ref. 1) and BLOSUM (Ref. 2) are the most popular for protein sequence comparisons]. Additionally, scores are chosen for **gaps**, which consist of one or more adjacent nulls in one sequence aligned with letters in the other. Because a single mutational event can insert or delete more than one residue, a long gap should be penalized only slightly more than a short gap. Accordingly, **affine gap costs**, which charge a relatively large penalty for the existence of a gap, and a smaller penalty for each residue it contains, have become the most widely used gap scoring system.

The practical effectiveness of sequence comparison depends critically upon the choice of appropriate substitution and **gap scores**. For **ungapped** local alignments, a complete theory exists describing which substitution scores best distinguish alignments representing true biological relationships from chance similarities. In brief, the score for aligning a given pair of residues i and j depends on the fraction q_{ij} of 'true alignment' positions in which these paired residues tend to appear[3]. Thus, defining a good substitution matrix comes down to estimating the **target frequencies** q_{ij} accurately.

After some thought, it is apparent that the desired target frequencies depend upon the degree of evolution divergence between the related sequences of interest. Thus, what is really required is not a single matrix, but rather a series of matrices tailored to varying degrees of evolutionary divergence[1–3]. This is precisely the perception underlying the construction of the PAM and BLOSUM series of amino acid substitution matrices. These matrices are generally used unmodified for **gapped** local and global alignment. There is no widely accepted theory for selecting gap costs, and their choice has generally been guided by trial and error[4].

0167-7799/98/$ – see front matter. Published by Elsevier Science.

```
(a)
P00001     1 MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNK---GI   58
               D  KG+ +F     QC T   +  K+  GP L G+ GRK G A G++Y+  N N    G+
P00090     1 Q-DAAKGEAVF----KQCMTCHRADKNMVGPALGGVVGRKAGTAAGFTYSPLNHNSGEAGL   56

P00001    59 IWGEDTLMEYLENPKKYIP--------------GTKMIFVGIKKKEERADLIAYLKKATNE  105
             +W ++ ++ YL +P  Y+                TKM F  +   ++R D+ AYL  AT +
P00090    57 VWTQENIIAYLPDPNAYLKKFLTDKGQADKATGSTKMTF-KLANDQQRKDVAAYL--ATLK  114

(b)
P13569  1221    EGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLRLL-----NTEGEIQIDGVS 1273
                +    ++  +S ++  G+ + L+G +GSGKS   +A L +L      T GEI  DG
P33593    13    QAAQPLVHGVSLTLQRGRVLALVGGSGSGKSLTCAATLGILPAGVRQTAGEILADGKP   70

P13569  1274    WDSITL---------QQWRKAFGVIPQKVFIFSGTFRKNLDPYEQWSDQEIWKVADEV 1322
                     L         Q  R AF  +              +  + + +   + K AD+
P33593    71    VSPCALRGIKIATIMQNPRSAFNPL------------HTMHTHARETCLALGKPADDA  116

P13569  1323    GLRSVIEQFP-GKLDFVLVDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDPV 1379
                 L + IE         VL      +S G  Q M +A +VL ++  ++ DEP+  LD V
P33593   117    TLTAAIEAVGLENAARVLKLYPFEMSGGMLQRMMIAMAVLCESPFIIADEPTTDLDVV  174
```

Fig. 1. Two protein alignments. **(a)** An optimal global alignment of human cytochrome c (105 residues; SWISS-PROT accession number P00001) and *Rhodopseudomonas palustris* cytochrome c_2 (114 residues; SWISS-PROT accession number P00090). **(b)** An optimal local alignment of the human cystic fibrosis transmembrane conductance regulator (1480 residues; SWISS-PROT accession number P13569) and *Escherichia coli* nickel transport ATP-binding protein NIKD (253 residues; SWISS-PROT accession number P33593). Scores for both alignments are calculated using the BLOSUM62 amino acid substitution matrix[2], and affine gap costs[7] that assign the score $-(11 + k)$ to a gap of length k. The global alignment, with score 131, is required to include the whole of the two input sequences and is constructed using the Needleman–Wunsch algorithm[5]. The local alignment, with score 89, involves only those segments of the two input sequences that optimize the score and is constructed using the Smith–Waterman algorithm[6]. On the central line of each alignment, characters indicate identical amino acids and '+' signs indicate similar amino acids (i.e. those whose alignment receives a positive score).

Alignment algorithms and database searches

After defining the score of an arbitrary alignment, one is faced with finding the **optimal** (i.e. highest scoring) alignment, or alignments, of two sequences. Fortunately, given additive scores as discussed above, a set of relatively efficient **dynamic programing** algorithms is available for this task. The first described in the biological literature was the **Needleman–Wunsch algorithm** for global alignment[5]. Subsequently, a slight variant was proposed, termed the **Smith–Waterman algorithm**, which can find the optimal local alignment of two sequences[6]. Both these algorithms require time proportional to the product of the lengths of the sequences being compared. Originally, neither could deal with affine gap costs, but both can now be modified to do so with only a small constant-factor decrease in speed[7].

Because similarities between DNA and protein sequences often span only segments of the sequences involved, the most popular database similarity search programs are based on the Smith–Waterman local alignment algorithm[6]. However, without special-purpose hardware or massively parallel machines the time required by Smith–Waterman renders it too slow for most users. The **FASTA** (http://www2.ebi.ac.uk/fasta3/) (Ref. 8) and **BLAST** (http://www.ncbi.nlm.nih.gov/BLAST) (Refs 9–11) programs therefore use **heuristic** strategies to concentrate their efforts on the sequence regions most likely to be related. Rapid exact-match procedures first identify promising regions, and only then is Smith–Waterman invoked. This approach permits FASTA and BLAST to run 10–100 times faster than full-blown Smith–Waterman, at the cost of overlooking an occasional similarity.

Some of the adjustable parameters of FASTA and BLAST control the details of their heuristics and thus influence the trade-off between speed and sensitivity. The effectiveness of any alignment program depends upon the scoring systems it employs[2–4]. Most importantly, protein similarities corresponding to true homologies are almost always easier to distinguish from chance than their corresponding DNA similarities, so coding DNA should always be conceptually translated to protein before performing a search. The practical use of database search programs is discussed in the article by Steven Brenner on pp. 9–12.

The statistics of alignment scores

To test the biological relevance of a global or local alignment of two sequences, one needs to know how great an alignment score can be expected to occur by chance. In this context, 'chance' can mean the comparison of: (1) real but unrelated sequences; (2) real sequences that are shuffled to preserve compositional properties; or (3) sequences that are generated randomly based upon a DNA or protein sequence model.

Very little of practical value is known about the random distribution of global alignment scores. One of the few ways to evaluate the significance of such a score is to generate an empirical score distribution from the alignment of many 'random' sequences of the same lengths as the two sequences being compared[12]. From this distribution, the ***Z* value** (the number of standard deviations from the mean) for the alignment score of interest can then be estimated. Importantly, it should not be assumed that the score distribution is normal; indeed, its general form is unknown. Therefore, an accurate significance estimate cannot currently be derived from the *Z* value.

Fortunately, much more is known about the statistics of local alignment scores. Under reasonable assumptions, the random score distribution for optimal ungapped local alignments can be proved to follow an **extreme value distribution**[13,14]. Such a proof is unavailable for gapped local alignments, but computational experiments strongly suggest that the same type of distribution applies[10]. An essential property of the extreme value distribution is that its right-hand tail decays exponentially in x, as opposed to x^2 for the normal distribution. Improperly assuming a normal distribution for optimal local alignment scores can thus result in gross exaggerations of statistical significance.

Current versions of the FASTA and BLAST search programs report the **raw scores** of the alignments they return, as well as assessments of their statistical significance, based upon the extreme value distribution. Most simply, these assessments take the form of ***E* values**. The *E* value for a given alignment depends upon its score, as well as the lengths of both the query sequence and the database searched. It represents the number of distinct alignments with equivalent or superior score that might have been expected to have occurred purely by chance. Thus an *E* value of five is not statistically significant, whereas an *E* value of 0.01 is. BLAST also reports **bit scores**, which are scaled versions of the raw scores[11]. A bit score takes into account the statistical parameters[3,10,13] of the scoring system employed, and is therefore more informative than a raw score for describing the quality of an alignment.

Masking regions of restricted composition

Many DNA and protein sequences contain regions of highly restricted nucleic acid and amino acid composition and regions of short elements repeated many times[15]. The standard alignment models and scoring systems were not designed to capture the evolutionary processes that led to these **low-complexity regions**. As a result, two sequences containing compositionally biased regions can receive a very high similarity score that reflects this bias alone. For many purposes, these regions are uninteresting and can obscure other important similarities. Therefore, programs that filter low-complexity regions from query or database sequences will often turn a useless database search into a valuable one[15].

Multiple sequences

Global and local pairwise sequence comparison and alignment can be generalized to multiple sequences. From multiple alignments, **profiles** [related to hidden Markov models (**HMMs**)] can be abstracted and these can greatly enhance the sensitivity of database search methods to evolutionarily distant and subtle sequence relationships[11]. As discussed by Sean Eddy on pp. 15–18 and by Kay Hofmann on pp. 18–21, this area is increasingly becoming the focus of algorithm and database development for biological sequence comparison.

Dedication

This article is dedicated to Dr Bruce W. Erickson, friend and mentor.

References

1 Dayhoff, M.O. *et al.* (1978) in *Atlas of Protein Sequence and Structure* (Vol. 5, Suppl. 3) (Dayhoff, M.O., ed.), pp. 345–352, National Biomedical Research Foundation
2 Henikoff, S. and Henikoff, J.G. (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 10915–10919
3 Altschul, S.F. (1991) *J. Mol. Biol.* 219, 555–565
4 Pearson, W.R. (1995) *Protein Sci.* 4, 1145–1160
5 Needleman, S.B. and Wunsch, C.D. (1970) *J. Mol. Biol.* 48, 443–453
6 Smith, T.F. and Waterman, M.S. (1981) *J. Mol. Biol.* 147, 195–197
7 Gotoh, O. (1982) *J. Mol. Biol.* 162, 705–708
8 Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448
9 Altschul, S.F. *et al.* (1990) *J. Mol. Biol.* 215, 403–410
10 Altschul, S.F. and Gish, W. (1996) *Methods Enzymol.* 266, 460–480
11 Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
12 Fitch, W.M. (1983) *J. Mol. Biol.* 163, 171–176
13 Karlin, S. and Altschul, S.F. (1990) *Proc. Natl. Acad. Sci. U. S. A.* 87, 2264–2268
14 Dembo, A. *et al.* (1994) *Ann. Probab.* 22, 2022–2039
15 Wootton, J.C. and Federhen, S. (1993) *Comput. Chem.* 17, 149–163

Practical database searching

Sequence comparisons need to be performed as carefully as wet-lab procedures, in terms of both experimental design and interpretation. The basic requirements of database searching, the factors that can affect the search results and, finally, how to interpret the results are discussed.

Steven E. Brenner
Dept of Structural Biology,
Stanford University, Stanford,
CA 94305-5400, USA.
brenner@hyper.stanford.edu

More sequences have been putatively characterized by database searches than by any other single technology. For good reason: programs like **BLAST** are fast and reliable. However, sequence comparison procedures should be treated as experiments analogous to standard laboratory procedures. Their use deserves the same care both in the design of the experiment and in the interpretation of results.

The database search experiment

Design of a BLAST database search requires consideration of what information is to be gained about the query sequence of interest. The main constraint is that database searching can only reveal similarity. However, from this similarity, homology (i.e. evolutionary relationship) can be inferred and, from that, one might be able to infer function. Although the former inference is now reliable

Box 1. Database searching: basic considerations

- Think about every step
- Search a large current database
- Compare as protein rather than DNA
- Filter query for low-complexity regions
- Interpret scores with *E* values
- Recognize that most homologs are not found by pairwise sequence comparison
- Consider slower and more powerful methods, but use iterative programs with great caution

for carefully performed sequence comparison, the second is still fraught with challenges. Box 1 provides some guidelines for performing reliable and sensitive database searches.

Planning a good experiment requires an understanding of the method being applied. Fundamentally, database searches are a simple operation: a query sequence is locally aligned with each of the sequences (called targets) in a database. Most programs, such as BLAST (Ref. 1) and **FASTA** (Ref. 2), use **heuristics** to speed up the alignment procedure, while the **Smith–Waterman algorithm**[3] (implemented, for example, in **SSEARCH**) rigorously compares the query sequence with each target in the database.

A score is computed from each alignment, and the query–target pairs with the best scores are then reported to the user. Typically, statistics are used to help improve the interpretation of these scores. A more detailed description of the process can be found in the article by Stephen Altschul on pp. 7–9. Although BLAST is the most widely used tool for sequence comparison, many other programs can help identify, confirm and interpret distant evolutionary relationships.

Databases, programs and comparison types

Formulation of the experiment begins with a decision about what types of sequences to compare: DNA, protein or DNA as protein. If the sequence under consideration is a protein or codes for a protein, then the search should probably take place at the protein level, because proteins allow one to detect far more distant homology than does DNA[2,4]. For example, in DNA comparisons, there is noise from the rapidly mutated third-base position in each codon and from comparisons of noncoding frames (although this latter issue still arises in DNA-as-protein searches). In addition, amino acids have chemical characteristics that allow degrees of similarity to be assessed rather than simple recognition of identity or non-identity. For these reasons, DNA versus DNA comparison (using the blastn program) is typically only used to find identical regions of sequence in a database. One would carry out such a search to discover whether the gene has been previously sequenced and to determine where it is expressed or where splice junctions occur. In short, protein-level searches are valuable for detecting evolutionarily related genes, while DNA searches are best for locating nearly identical regions of sequence.

Next, it is necessary to select a database to search against. For homology searches, the most commonly searched database on the NCBI (National Center for Biotechnology Information) website is the **nr database**. The nr protein database combines data from several sources, removes the redundant identical sequences and yields a collection with nearly all known proteins. The NCBI nr database is frequently updated in order to incorporate as many sequences as possible. Obviously, a search will not identify a sequence that has not been included in the database and, as databases are growing so rapidly, it is essential to use a current database. Several specialized databases are also available, each of which is a subset of the nr database. ***E*-value** statistics (discussed below) are affected by database size, so, if you are interested in searching for proteins of known structure, it is best to just search the smaller **pdb database**.

One might also wish to search DNA databases at the protein level. Programs can do so automatically by first translating the DNA in all six reading frames and then making comparisons with each of these conceptual translations. The nr DNA database, which contains most known DNA sequences except **GSS**s, **EST**s, **STS**s and **HTGS**s, is useful to search when hunting new genes; the identified genes in this database would already be in the protein nr database. Searches against the GSS, EST, STS and HTGS databases can find new homologous genes and are especially useful for learning about expression data or genome map location.

Because of the different combinations of queries and database types, there are several variants of BLAST (see Table 1). Note that it is desirable to use the newest versions of BLAST, which support **gapped alignments** (see the article by Stephen Altschul on pp. 7–9). The older versions are slower, detect fewer homologs and have problems with some statistics. The programs can be run over the World Wide Web (WWW) and can be downloaded from an **ftp** site to run locally. Another option is to use the FASTA package[2]. The FASTA program can be slower but more effective than BLAST. The package also contains SSEARCH, an implementation of the rigorous Smith–Waterman algorithm, which is slow but the most sensitive. As described in the article by Sean Eddy on pp. 15–18, iterative programs such as **PSI-BLAST** require extreme care in their operation because they can provide very misleading results; however, they have the potential to find more homologs than purely pairwise methods.

Filtering

The statistics for database searches assume that unrelated sequences will look essentially random with respect to each other. However, certain patterns in sequences violate this rule. The most common exceptions are long runs of a small number of different residues (such as a polyalanine tract). Such regions of sequence could spuriously obtain extremely high match scores. For this reason, the NCBI BLAST server will automatically remove such sections in proteins (replacing them with an X) using the **SEG** program[5] if 'default **filtering**' is selected. DNA sequences will be similarly **masked** by **DUST**.

Table 1. BLAST variants for different searches[a]

Program	Query	Database	Comparison	Common use
blastn	DNA	DNA	DNA level	Seek identical DNA sequences and splicing patterns
blastp	Protein	Protein	Protein level	Find homologous proteins
blastx	DNA	Protein	Protein level	Analyze new DNA to find genes and seek homologous proteins
tblastn	Protein	DNA	Protein level	Search for genes in unannotated DNA
tblastx	DNA	DNA	Protein level	Discover gene structure

[a] Similar variant programs are available for FASTA. Protein-level searches of DNA sequences are performed by comparing translations of all six reading frames.

Although these programs automatically remove the majority of problematic matches, some problems invariably slip through; moreover, valid hits might be missed if part of the sequence is masked. Therefore, it might be helpful to try using different masking parameters.

Other sorts of filtering are also often desirable. For example, **iterative searches** are prone to contamination by regions of proteins that resemble coiled coils or transmembrane helices. The problem is that a protein that is similar only in these general characteristics might match initially. The profile then emphasizes these inappropriate characteristics, eventually causing many spurious hits. Heavily cysteine-rich proteins can also obtain anomalous high scores. Especially if these characteristics are not filtered, it is necessary to review the alignment results carefully to ensure that they have not led to incorrect matches.

Alignment, algorithmic and output parameters

Three other sets of parameters also affect search results, but they rarely require careful consideration by most users. First, the matrix and gap parameters determine how similarity between two sequences is determined. When two residues in a protein are aligned, programs use the matrix to determine whether the amino acids are similar (and thus receive a positive score) or very different. The default matrix for BLAST is called BLOSUM62 (Ref. 6), and the programs will not currently operate reliably with other matrices. The gap parameters determine how much an alignment is penalized for having gaps: the existence parameter is a fixed cost for having a gap and the per-position cost is a cost dependent upon the length (i.e. the number of residues). Typically, there is a large cost associated with introducing a gap and a small additional cost such that longer gaps are worse. It is rarely very beneficial to change these from their defaults.

The second set of parameters determines the heuristics that BLAST uses. By altering these numbers, it is possible to make the program run slower and be more sensitive, or to run faster at the cost of missing more homologs. The complexity of these parameters in BLAST precludes extensive description here. Currently, it is very rare for users to alter these options from the defaults. The FASTA program has one such parameter, called **ktup**, that a user will often want to set. Searches with ktup = 1 are slower, but more sensitive, than BLAST; ktup = 2 is fast, but less effective.

A third set of parameters regulates how many results are reported. By default, the programs will report only matches with an ***E* value** (described below) up to 10. The total number of matches is limited to the best 500, and detailed information with the alignment is provided for up to 100 pairs. To retrieve more matches, these numbers can be increased.

Interpretation of results

Interpretation of the results of a sequence database search involves first evaluating the matches to determine whether they are significant and therefore imply homology. The most effective way of doing this is through use of statistical scores or *E* values. The *E* values are more useful than the **raw** or **bit scores**, and they are far more powerful than percentage identity (which is best not even considered unless the identity is very high)[7]. Fortunately, the *E* values from FASTA, SSEARCH and NCBI gapped BLAST seem to be accurate and are therefore easy to interpret (see Ref. 7).

The *E* value of a match should measure the expected number of sequences in the database that would achieve a given score by chance. Therefore, in the average database search, one expects to find ten random matches with *E* values below 10; obviously, such matches are not significant. However, lacking better matches, sequences with these scores might provide hints of function or suggest new experiments. Scores below 0.01 would occur by chance only very rarely and are therefore likely to indicate homology, unless biased in some way. Scores of near 1e−50 (1×10^{-50}) are now seen frequently, and these offer extremely high confidence that the query protein is evolutionarily related to the matched target in the database.

Inferring function from the homologous matched sequences is a problematic process. If the score is extremely good and the alignment covers the whole of both proteins, then there is a good chance that they will share the same or a related function. However, it is dangerous to place too much trust in the query having the same function as the matched protein; functions do diverge, and organismal or cellular roles can alter even when biochemical function is unchanged. Moreover, a significant fraction of functional annotations in databases are wrong, so one needs to be careful. There are other complexities; for example, if only a portion of the proteins align, they might share a domain that only contributes one aspect of the overall function. It is often the case that all of the highest-scoring hits align to one region of the query, and matches to other regions need to be sought much lower

URLs

BLAST Web site
http://www.ncbi.nlm.nih.gov/BLAST/

BLAST FTP site
ftp://ncbi.nlm.nih.gov/blast/

FASTA at EBI
http://www2.ebi.ac.uk/fasta3/

FASTA FTP site
ftp://ftp.virginia.edu/pub/fasta

Sequence search site
http://sss.stanford.edu/sss/

in the score ranking. For this reason, it is necessary to consider carefully the overlap between the query and each of the targets.

Database search methods are also limited because most homologous sequences have diverged too far to be detected by pairwise sequence comparison methods[7]. Thus, failure to find a significant match does not indicate that no homologs exist in the database; rather, it suggests that either more-powerful computational methods (such as those described by Sean Eddy on pp. 15–18 and by Kay Hofmann on pp. 18–21) or experiments would be necessary to locate them.

Conclusion

One should neither have excessive faith in the results of a BLAST run nor blithely disregard them. The BLAST programs are well-tested and reliable indicators of sequence similarity, and their underlying principles are straightforward. Complexities added by the fast algorithms typically need not be carefully considered, because the program and its parameters have been optimized for hundreds of thousands of daily runs. If one is careful about posing the database search experiment and interprets the results with care, sequence comparison methods can be trusted to provide an incomparable wealth of biological information rapidly and easily.

References

1 Altshul, S.F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402

2 Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448

3 Smith, T.F. and Waterman, M.S. (1981) *J. Mol. Biol.* 147, 195–197

4 States, D.J. *et al.* (1991) *Methods* 3, 66–70

5 Wootton, J.C. (1994) *Comput. Chem.* 18, 269–285

6 Henikoff, S. and Henikoff, J.G. (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 10915–10919

7 Brenner, S.E., Chothia, C. and Hubbard, T.J.P. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 6073–6078