Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 5873–5877, June 1993
Evolution

Applications and statistics for multiple high-scoring segments in molecular sequences

(sequence comparison/pattern recognition/molecular features/statistical significance)

SAMUEL KARLIN† AND STEPHEN F. ALTSCHUL‡§

†Department of Mathematics, Stanford University, Stanford, CA 94305; and ‡National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

Contributed by Samuel Karlin, March 24, 1993

ABSTRACT **Score-based measures of molecular-sequence features provide versatile aids for the study of proteins and DNA. They are used by many sequence data base search programs, as well as for identifying distinctive properties of single sequences. For any such measure, it is important to know what can be expected to occur purely by chance. The statistical distribution of high-scoring segments has been described elsewhere. However, molecular sequences will frequently yield several high-scoring segments for which some combined assessment is in order. This paper describes the statistical distribution for the sum of the scores of multiple high-scoring segments and illustrates its application to the identification of possible transmembrane segments and the evaluation of sequence similarity.**

The study of molecular-sequence data can be assisted by statistical methods of sequence analysis. Among the aims of such study is the discovery of patterns relevant to genomic organization, nucleic acid processing, protein folding, and biochemical function as well as their evolutionary developments. A region of unusual amino acid composition in a protein sequence may correlate with a specific biological function. Similarly, the conservation over evolutionary time of segments shared by different proteins may provide clues to structure and function.

Among the tools for detecting interesting regions in protein sequences are score-based methods. These assign appropriate positive numerical values to amino acids likely to be found within the type of region sought and negative values to residues unlikely to occur. Since scores permit differentiation between residues, they engender more sensitive analyses than do measures that consider only simple matching. Scores have been used to locate transmembrane or significantly hydrophobic segments, DNA-binding domains, and regions of concentrated charge (1). They are also employed to identify similar regions shared by two or more protein or DNA molecules and are at the core of many sequence data base search programs (2–4).

A crucial question for any given score-based or other measure applied to molecular sequences is what can be expected to occur purely by chance. Empirical statistical studies can be based upon sequence data collections (1, 5–8) or upon permutations of sample sequences (9, 10). In addition, analytic statistical results can afford calculable criteria for the evaluation of sequences and can elucidate the function of the parameters in the measures to which they apply (11). They provide means for recognizing outliers, for developing contrasting sequence classifications, and for comparing different data sets in a consistent manner.

The greatest limitation on the analytic approach is the difficulty of deriving statistical distributions for any but the simplest sequence measures. Among those that have yielded to analysis are ones based on runs of a given residue type, allowing for a specified number or proportion of mismatches (12–14). More recently, a theory has been developed for characterizing unusual sequence patterns, defined with reference to general scoring systems (15–18). Scores may be based on residue biochemical or physical properties or, in the case of sequence comparison, on residue similarities. Specifically, the theory describes the asymptotic extremal distribution of high aggregate segment scores as well as the letter composition of high-scoring segments.

In this paper we consider several natural extensions of score-based measures. An important such extension is the sum of the r greatest segment scores. This measure is appropriate when there may be several distinct segments of a given type within a protein or DNA sequence (e.g., transmembrane segments). Also, for sequence comparisons, the existence of insertions or deletions can break an alignment into several pieces, and the sum of their scores can be an appropriate measure of local sequence similarity. From this consideration there arises the problem of "consistency" for high-scoring segment pairs: the requirement that multiple pairs be combinable into a single "gapped" alignment. We discuss below how this constraint affects the distribution of the sum statistic. The use of these statistics will be illustrated with several examples.

The Statistical Theory for High-Scoring Segments

Given a molecular sequence, we assume that scores are assigned to the various sequence elements and study the statistical behavior of the segment (of whatever length) with greatest aggregate score. In this section we review briefly the theory for such maximal-segment scores (15–18). The basic themes of this theory are visible in the various extensions that follow.

In the simple "independence" random sequence model we employ, the elements of a sequence are chosen independently from an alphabet of a letters with respective probabilities $p_1, \ldots, p_a$. A DNA sequence, for example, would have $a = 4$, and a protein sequence using the standard alphabet would have $a = 20$. Theory exists for the more complicated case of Markov-dependent sequences but will not be discussed here (17). A score s_i is assigned to each type of letter. For proteins these scores may be based, for example, on physicochemical or structure-related properties such as charge, size, hydrophobicity, and helix-forming potential. The maximal segment of a sequence is defined as that contiguous string of letters with greatest aggregate score. The random distribution for the score S of this segment can be expressed by using three parameters, which are described below.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: p.d.f., probability density function; PIR, Protein Identification Resource.

§To whom reprint requests should be addressed.

A necessary assumption for the following theory is that the expected score per letter $\Sigma_i p_i s_i$ be negative. [Scores based on likelihood ratios (15) always satisfy this condition.] Were the expected score positive, the maximal segment would always tend to be virtually the entire sequence, so such a scoring system would not be of much use for identifying unusual regions. The existence of at least one positive score along with the previous condition implies that there is always a unique positive solution $x = \lambda$ to the equation $\Sigma_i p_i e^{s_i x} = 1$. The parameter λ may be thought of as a natural scale for the scoring system employed.

A second parameter K, for which an explicit but more complex formula is available, is also readily calculated from the scores s_i and their background probabilities p_i (15–18). The final relevant parameter is the length N of the random sequence from which the maximal segment is drawn. The statistical theory is then most simply expressed in terms of the normalized score $S' = \lambda S - \ln KN$. For large N, the tail probability (Prob) that S' is greater than or equal to x is well approximated by the formula

$$\text{Prob}(S' \geq x) \approx 1 - \exp(-e^{-x}). \quad [1]$$

For sequence comparison, the theory has a parallel development. Scores s_{ij} are now assigned not to individual letters but to pairs of letters. Given two sequences, the maximal-segment pair is simply that pair of equal-length segments, one from each sequence, which when aligned have maximal aggregate score S. The expected score per residue pair must still be negative, and the formulas for λ and K are the same as before. The main difference is that the "search space size" parameter N becomes the product of the lengths of the two sequences being compared. A number of conditions must hold for $\text{Prob}(S' \geq x)$ to converge to formula 1 for large N (A. Dembo, S.K., and O. Zeitouni, unpublished data), but this formula is always conservative—i.e., it provides an upper bound on the desired probability.

For single-sequence protein analysis, scores appropriate for the detection of transmembrane segments and DNA-binding domains have been described (1, 6, 19). For sequence comparison, a wide range of scoring systems have been proposed (11, 20–28), and segment-pair scores underpin the BLAST data base search programs (2, 29, 30). These are examples where the basic theory finds direct application. In many cases, however, more sophisticated scoring methods, such as those studied in this paper, are appropriate.

The Statistical Theory for Multiple High-Scoring Segments

Sometimes a single molecule will contain multiple regions with a common property of biological interest, or two molecules will share several quite similar regions. For example, a protein may have several distinct transmembrane segments or regions of concentrated charge. Two proteins may share a number of regions of conserved secondary structure, separated by loops of variable length and composition. When this is the case, seeking the single highest-scoring region can discard much valuable information. We therefore consider the scores $S_1, \ldots, S_r$ of the r highest-scoring distinct segments. Statistics for these random variables are most conveniently written using the normalized scores $S'_k = \lambda S_k - \ln KN$. For large N, the joint probability density function (p.d.f.) for $S'_1, \ldots, S'_r$ is well approximated by the formula

$$f(x_1, \ldots, x_r) = \exp\left(-e^{-x_r} - \sum_{k=1}^{r} x_k\right), \quad [2]$$

where $x_1 \geq x_2 \geq \cdots \geq x_r$. The distribution of any function of the S'_k may be calculated from this distribution. The simplest application is to calculate the tail probability that S'_r is greater than or equal to x; integration yields

$$\text{Prob}(S'_r \geq x) \approx 1 - \exp(-e^{-x}) \sum_{k=0}^{r-1} \frac{e^{-kx}}{k!}, \quad [3]$$

which is a generalization of formula 1.

Of greater interest and utility is the distribution of the sum of the r highest normalized scores $T_r = S'_1 + \cdots + S'_r$. From formula 2 and some algebraic manipulation, one may show that for large N, the p.d.f. for T_r approaches

$$f(t) = \frac{e^{-t}}{r!(r-2)!} \int_0^\infty y^{r-2} \exp(-e^{(y-t)/r})\, dy. \quad [4]$$

All moments of this distribution may be calculated by means of Laplace transforms. The mean is given by $r(1 + \gamma - \Sigma_{k=1}^{r} 1/k)$, where $\gamma \approx 0.577$ is Euler's constant, and is well approximated by $r(1 - \ln r) - \frac{1}{2}$. The variance is $r^2(\pi^2/6 - \Sigma_{k=1}^{r} 1/k^2) + r$, which is approximately $2r - \frac{1}{2}$.

To obtain the tail probability that $T_r \geq x$, one must integrate Eq. 4 for t from x to infinity. This double integral is easily calculated numerically, and a program for the purpose in the C programming language is available from the authors. In the limit of large x, this tail probability behaves as

$$\text{Prob}(T_r \geq x) \approx \frac{e^{-x} x^{r-1}}{r!(r-1)!}. \quad [5]$$

Applications of these results will be given below.

Consistently Ordered Segment Pairs in Sequence Alignments

Two proteins may share distinct, homologous domains that need not retain the same relative order. More often, however, separate high-scoring segment pairs arise from insertions or deletions within a matching region. In the context of pairwise sequence comparison, one may wish to exclude the former possibility and consider only the latter; this simultaneously excludes from analysis cases that do not fit the biological model and increases the statistical significance of those segment pair sets that do.

Requiring that a collection of high-scoring segment pairs be consistent with a single alignment including gaps imposes a certain "geometry" on the pairs that so far has not been taken into account. For a given high-scoring segment pair i, let (x_i, y_i) indicate the midpoints of its constituent segments within their respective sequences. A necessary condition for combining several segment pairs into a single consistent alignment is that for any two pairs i and j, $x_i < x_j$ if and only if $y_i < y_j$. We will call a set of segment pairs "consistently ordered" if it satisfies this condition.

The random variable T_r from the previous section can be written as $\lambda(\Sigma_{k=1}^{r} S_k) - \ln K^r - \ln N^r$. The last term may be understood as correcting for the N^r different possible sets of starting positions for the r segment pairs whose scores are the S_k. (Remember that for pairwise sequence comparison, N is the product of the lengths of the sequences being compared.) If we require a set of r segment pairs to be consistently ordered before allowing their scores to be combined, we effectively divide the size of the possible solution space by $r!$. Therefore, if T^*_r is the greatest value attainable as the sum of normalized scores S'_k from r distinct and consistently ordered segment pairs, $T^*_r + \ln r!$ has a p.d.f. approaching that of Eq. 4 for large N.

This analysis can be extended to more restrictive constraints on the relationship of combined segment pairs. Between pairs, for example, one may allow gaps within each

Table 1. High-scoring segments of the *D. virilis* sevenless protein (34) [Protein Identification Resource (PIR) code A35774] with their associated scores and *P* values

Segment	Positions	Score (normalized score)	*P* value	
			Segment	Sum
LVLAIIAPAAIVSSCVLALVLV	2141–2162	67 (4.4)	0.012	0.012
FLVTGHGGISTILIANLLLLLLLSL	116–140	55 (2.5)	0.080	0.0035
ISAPIIVALLAL	466–477	38 (−0.2)	0.71	0.0053

Segment scores are based on the transmembrane scores from ref. 1.

sequence of only some maximum size. The appropriate statistics then depend upon the extent to which the space of possible solutions is reduced.

The measure presented in this section combines gaps and scores in a natural manner. One drawback of this measure, however, is that so long as a gap is permitted at all, no premium is placed on a short as opposed to a long one. A statistical problem that remains open is the random distribution of scores from optimal alignments that include gaps and for which length-dependent gap costs are assessed (4, 31). While numerical studies have been conducted on the statistics of such scoring systems (7, 8), they have resisted complete analysis to date.

Further Results Involving Consistent Ordering

One question similar in spirit but different in detail from those considered above is how many distinct segments can be expected with score at least x. This question is most easily answered by using the composite parameter $y = KNe^{-\lambda x}$. For large N and for x sufficiently large that y is not much greater than 1, the number of distinct segments whose score is at least x is then approximately Poisson distributed with parameter y (15–18). In other words, the probability of observing exactly k such segments is approximately $e^{-y}y^k/k!$. The probability of observing at least r segments with score at least x is calculated by summing this quantity for k from r to infinity.

In the case of sequence alignments, we now wish to impose the additional requirement of consistent ordering. The most natural question concerns the probability that there are at least r distinct and consistently ordered segment pairs all with score at least x. The desired probability arises if each term of the infinite sum is multiplied by the probability that a set of k segment pairs contains a consistently ordered subset of size at least r. For large N this probability can be seen to approach $R_{k,r}/k!$, where $R_{k,r}$ is the number of permutations of the integers 1 to k that contain an increasing subsequence of length at least r. Thus, the formula for the desired probability becomes

$$e^{-y}\sum_{k=r}^{\infty}\frac{y^kR_{k,r}}{k!^2}. \qquad [6]$$

To employ formula 6 effectively, one must be able to calculate $R_{k,r}$ for at least the first several k values greater than or equal to r. When $r \leq 4$, general formulas are available for $R_{k,r}$ (32). Moreover, for all r, various combinatorial facts about permutations (33) suffice to prove that $R_{r,r} = 1$; $R_{r+1,r} = r^2 + 1$, and $R_{r+2,r} = (r^4 + 2r^3 + r^2 + 2r + 6)/2$. Specific but increasingly complicated formulas may be derived for successive terms. However, the first three terms just given should be sufficient for most purposes.

Examples

To illustrate the use of sum statistics, we first consider the sevenless protein from the fruit fly *Drosophila virilis* (34). This molecule is a tyrosine kinase receptor required for embryogenesis of the eye; it is known to have one and is suspected to have two transmembrane domains (35). We analyzed the molecule for transmembrane segments, using scores derived for this purpose by Karlin and Brendel (1). The three highest-scoring segments of the protein are shown in Table 1, arranged in decreasing order of score. For this analysis, the relevant statistical parameters may be calculated as $\lambda = 0.159$, $K = 0.21$, and $N = 2594$. The single highest-scoring segment, consisting of residues 2141–2162, has a normalized score of 4.4, which by formula 1 corresponds to a *P* value of 0.012. The second highest normalized score (for residues 116–140) is 2.5, corresponding to a *P* value of 0.08. Neither of these segments in isolation may be considered significant at the 99% level. However, as shown in Table 1, when analyzed in unison, the *P* value for the sum of their scores drops to 0.0035. Successive high-scoring segments (i.e., those other than the top-scoring two) do not improve the overall result. The two segments identified as statistically significant by this method are the putative transmembrane domains described in the original paper (34).

As a second example, we analyze the human serotonin receptor (36) for transmembrane segments. This molecule is a member of the large family of guanine nucleotide-binding protein-coupled receptors, which generally contain seven transmembrane segments, accounting for roughly half of the complete protein. The large proportion of hydrophobic residues within these proteins render concentrations of such

Table 2. High-scoring segments of the human serotonin receptor (36) (PIR code S07343), with their associated scores and *P* values

Segment	Positions	Score (normalized score)	*P* value	
			Segment	Sum
VITSLLLGTLIFCAVLGNACVVAAIAL	(37) 37–63 (62)	66 (3.5)	0.031	0.031
LGIIMGTFILCWLPFFIVALVL	(345) 346–367 (366)	65 (3.4)	0.034	0.0036
ALISLTWLIGFLISI	(152) 154–168 (177)	46 (1.2)	0.26	0.0019
IYSTFGAFYIPLLLMLVL	(191) 196–213 (216)	41 (0.6)	0.42	0.0011
LIGSLAVTDLMVSVLVLPMAAL	(74) 74–95 (98)	38 (0.3)	0.53	0.00064
LFIALDVLCCTSSILHLCAIAL	(110) 111–132 (134)	31 (−0.5)	0.81	0.00056
LLGAII	(378) 379–384 (402)	26 (−1.1)	0.95	0.00061

Segment scores are based on the transmembrane scores from ref. 1. Next to the position numbers representing the extent of each high-scoring segment are given in parentheses those for the corresponding putative transmembrane segment as specified in SWISS-PROT (38).

Table 3. High-scoring segment pairs, with their associated scores and P values, from a comparison of the chicken gene X protein (39) (PIR code DXCH) and the fowlpox virus antithrombin III homolog (40) (PIR code WMVZF3)

Segment pair	Positions	Score (normalized score)	P value	
			Segment pair	Sum
VYLPQMKIEEKYNLTSVLMALGMTDLF	125–151			
YLP E L L G DLF		52 (7.6)	4.7×10^{-4}	4.7×10^{-4}
LYLPKFELEDDVDLKDALIHMGCNDLF	44–70			
SANLTGISSAESLKISQAVHGAFMELSEDGIEMAGST	154–190			
S L GIS L I E G E A T		49 (6.7)	1.2×10^{-3}	4.2×10^{-6}
SGELVGISDTKTLRIGNIRQKSVIKVDEYGTEAASVT	72–108			
RADHPFLFLIKHNPTNTIVYFGRY	206–229			
A PF FL T G		46 (5.8)	3.1×10^{-3}	5.9×10^{-8}
KANVPFMFLVADVQTKIPLFLGIF	123–146			

Segment pair scores are calculated using the PAM-120 scoring matrix (11, 20). Amino acid identities are echoed on the central line of each alignment.

amino acids more difficult to distinguish from chance (cf. ref. 37).

Using the same transmembrane scores as before (1), the seven best segments of the human serotonin receptor are shown in Table 2, arranged in decreasing order of score. Here the relevant parameters are $\lambda = 0.114$, $K = 0.14$, and $N = 421$. The single highest-scoring segment consists of residues 37–63 and has a normalized score of 3.5, corresponding by formula 1 to a P value of 0.031. Therefore, neither this nor any of the other high-scoring segments may be considered, in isolation, particularly surprising. When several of the highest-scoring segments are analyzed in unison, however, the situation changes. As shown in Table 2, P values for the sum of the r highest segment scores continue to drop until $r = 6$, at which point the cumulative normalized score of 8.4 has a probability less than 6×10^{-4} of having occurred by chance. Further segments do not improve the overall result. It should be noted that the statistics for sums of high segment scores described above are valid only in the limit of large N. For a protein as short as the one in this example, they are inaccurate for $r > 2$. Nevertheless, even the sum of just the two highest segment scores provides good evidence that one is dealing with a multisegment transmembrane protein. Applications of the sum statistic to long DNA sequences will be discussed elsewhere.

Finally, to illustrate the use and potential power of sum statistics applied to pairwise sequence comparison, we consider an analysis of the chicken gene X protein (39) and the fowlpox virus antithrombin III homolog (40). When compared by using the PAM-120 amino acid substitution matrix (11, 20), three high-scoring segment pairs emerge (Table 3). Given this scoring system and the amino acid frequencies of the two sequences, $\lambda = 0.314$, $K = 0.17$, and $N = 34336$, yielding corrected scores of 7.6, 6.7, and 5.8 for the three alignments. From formula 1, the associated P values for these alignments are all less than 0.004, which is generally considered significant. However, such similarities are frequently uncovered in protein data base searches, in which tens of thousands of pairwise comparisons are typically performed (2). In such a multitrial context, P values near 10^{-6} generally are necessary before statistical significance may be claimed. None of the individual alignments shown in Table 3 achieve such significance, and any single one of them could easily arise by chance in a search of current protein sequence data bases (38, 41). This is no longer the case, however, when the three segment pairs are considered together. The sum of their normalized scores is 20.1, which for $r = 3$ corresponds to a P value of 5.9×10^{-8}, easily significant even in the context of a large data base search. Notice as well that the three segment pairs shown in Table 3 are consistently ordered. (In fact, the three pairs are in almost perfect alignment.) When this is taken as an *a priori* requirement for invoking a sum, the P value drops even further, to 1.2×10^{-8}. Thus, the ability to calculate statistics for the combined scores of distinct segment pairs can greatly increase the sensitivity of sequence comparison tools.

S.F.A. thanks Dr. John Spouge for helpful conversations and Dr. Warren Gish for programming assistance. We appreciate valuable comments on the manuscript from Dr. Volker Brendel. S.K. was supported in part by National Institutes of Health Grants GM39907-02 and GM10452-26 and National Science Foundation Grant DMS86-06244.

1. Karlin, S. & Brendel, V. (1992) *Science* **257**, 39–49.
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
4. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
5. Collins, J. F., Coulson, A. F. W. & Lyall, A. (1988) *Comput. Appl. Biosci.* **4**, 67–71.
6. Karlin, S., Bucher, P., Brendel, V. & Altschul, S. F. (1991) *Annu. Rev. Biophys. Biophys. Chem.* **20**, 175–203.
7. Mott, R. (1992) *Bull. Math. Biol.* **54**, 59–75.
8. Smith, T. F., Waterman, M. S. & Burks, C. (1985) *Nucleic Acids Res.* **13**, 645–656.
9. Altschul, S. F. & Erickson, B. W. (1985) *Mol. Biol. Evol.* **2**, 526–538.
10. Fitch, W. M. (1983) *J. Mol. Biol.* **163**, 171–176.
11. Altschul, S. F. (1991) *J. Mol. Biol.* **219**, 555–565.
12. Arratia, R., Gordon, L. & Waterman, M. S. (1986) *Ann. Stat.* **14**, 971–993.
13. Arratia, R. & Waterman, M. S. (1989) *Ann. Probab.* **17**, 1152–1169.
14. Karlin, S. & Ost, F. (1988) *Ann. Probab.* **16**, 535–563.
15. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
16. Dembo, A. & Karlin, S. (1991) *Ann. Probab.* **19**, 1737–1755.
17. Karlin, S. & Dembo, A. (1992) *Adv. Appl. Probab.* **24**, 113–140.
18. Karlin, S., Dembo, A. & Kawabata, T. (1990) *Ann. Stat.* **18**, 571–581.
19. Brendel, V., Bucher, P., Nourbakhsh, I. R., Blaisdell, B. E. & Karlin, S. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 2002–2006.
20. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 3, pp. 345–352.
21. Feng, D. F., Johnson, M. S. & Doolittle, R. F. (1985) *J. Mol. Evol.* **21**, 112–125.
22. Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992) *Science* **256**, 1443–1445.
23. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.
24. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) *Comput. Appl. Biosci.* **8**, 275–282.

25. McLachlan, A. D. (1971) *J. Mol. Biol.* **61**, 409–424.
26. Schwartz, R. M. & Dayhoff, M. O. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 3, pp. 353–358.
27. States, D. J., Gish, W. & Altschul, S. F. (1991) *Methods* **3**, 66–70.
28. Wilbur, W. J. (1985) *Mol. Biol. Evol.* **2**, 434–447.
29. Altschul, S. F. & Lipman, D. J. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 5509–5513.
30. Gish, W. & States, D. J. (1993) *Nature Genet.* **3**, 266–272.
31. Sellers, P. H. (1984) *Bull. Math. Biol.* **46**, 501–514.
32. Gessel, I. M. (1990) *J. Combinat. Theory A* **53**, 257–285.
33. Knuth, D. E. (1973) *The Art of Computer Programming* (Addison–Wesley, Reading, MA), Vol. 3, pp. 48–72.
34. Michael, W. M., Bowtell, D. D. & Rubin, G. M. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 5351–5353.
35. Simon, M. A., Bowtell, D. D. & Rubin, G. M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 8333–8337.
36. Kobilka, B. K., Frielle, T., Collins, S., Yang-Feng, T., Kobilka, T. S., Francke, U., Lefkowitz, R. J. & Caron, M. G. (1987) *Nature (London)* **329**, 75–79.
37. Karlin, S., Brendel, V. & Bucher, P. (1992) *Mol. Biol. Evol.* **9**, 152–167.
38. Bairoch, A. & Boeckmann, B. (1992) *Nucleic Acids Res.* **20**, 2019–2022.
39. Heilig, R., Perrin, F., Gannon, F., Mandel, J. L. & Chambon, P. (1980) *Cell* **20**, 625–637.
40. Tomley, F., Binns, M., Campbell, J. & Boursnell, M. (1988) *J. Gen. Virol.* **69**, 1025–1040.
41. Barker, W. C., George, D. G., Mewes, H. W. & Tsugita, A. (1992) *Nucleic Acids Res.* **20**, 2023–2026.