Database Searching

1 Introduction

The set-up for a typical molecular Database search is as follows: Given are a query sequence and a database of sequences of the same type (Protein or DNA). For each entry of the database a local alignment to the query is calculated and the corresponding score is stored. The database entries are listed in decreasing order of their scores. Significantly high scores indicate evolutionary relationships (Homology).

In typical applications the query is a sequence where nothing is known about. The database search hopefully yields a set of homologous sequences. If there function is known and annotated in the database, we have a clue to the function of the query as well. To test the performance of database searches and the reliability of its results one can use members of well studied sequence families as queries. A typical result looks like this:

1. On top of the list are clear hits. These sequences are almost identical and their scores are far above average. Typical for such strong similarities are corresponding sequences in closely related species. (e.g. cytochrome-c of man and chimp).

2. Further down in the list we observe a twilight zone, consisting of both sequences that are distantly related to the query and sequences that are unrelated to it. However they all share about the same amount of similarity to the query (if measured in alignment score points).

3. Finally all the other unrelated sequences of the database are listed with small scores.

Let's come back to the application of analyzing sequences of unknown function. If we find a database entry that is "mostly identical" to the query (as those described in 1), we are done. These similarities do not occur just by chance and the only reasonable explanation left is common ancestry. The function of homologous sequences may still be different, hence further (wet) experiments are needed to determine their function reliably. Database searches are only supposed to yield homologous sequences to the query that guideline the further investigation of

its function. If we only get scores in the range described in 3 the search failed. Finally if we find similarities with score levels in the twilight we don't know. Homology can not be reliably inferred, however there is some reasonable chance that we found something interesting. The twilight zone is a superposition of distantly related sequences and not related sequences that score high just by chance. Natural questions are:

- Which score levels reveal clear hits?
- Which are ambiguous?
- Which are meaningless?

In statistical terms: We are interested in the significance of score levels. Significance calculations need a model for unrelated sequences (null hypothesis). We use independent pairs of i.i.d. sequences. The distribution of each position in the sequences is as an average distribution of nucleotides in DNA or amino acids in Proteins. Studies for Markov dependent sequences are also described in the literature Waterman and Vingron (1994). There is no model for related pairs of sequences in this theory, however statistical studies modeling both related and unrelated sequences are also described in the literature Hwa and Lässig (1998).

Database searching is challenging in the case of remote relationship. In this case similarity is normally restricted to conserved segments of both sequences. Even if both sequences are homologous in their entire length, it may occur that significant similarity only remains in segments that were subject to considerably strong evolutionary pressure. Hence database searching requires local comparison. The golden standard is to perform database searches by calculating local alignments using the Smith-Waterman algorithm. The more popular programs FASTA and BLAST are heuristic approximations of the Smith-Waterman algorithm that are sufficiently quick to search entire databases in minutes. BLAST does not use gaps, however it still performs good in database searches. Local comparison bears a characteristic statistical problem: Long sequences in the database provide more different pattern than short ones, hence they are more likely to score high just by chance. Hence we need to study the significance of score levels given the length of the sequences that where compared.

2 Notations

Let $\mathcal{A}$ be a finite alphabet, the 4 letter alphabet of DNA or the 20 letter alphabet of amino acids. Let $\mu = (\mu_1, ..., \mu_k)$ be a distribution vector on the alphabet. μ_i denoting the relative frequency of letter i. Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be i.i.d

sequences. Both X_i and Y_j are distributed according to μ and the two sequences are independent from each other. Further we are given a score function S assigning a real number to each pair of letters. These score functions are usually denoted as matrices (e.g. PAM 250 for amino acids). The expectation of S under the null hypothesis should be negative, i.e.

$$E[S] = \sum_{a,b} \mu_a \mu_b \, S_{ab} = m \; < \; 0.$$

The optimal alignment score without gaps of the first n respectively m variables is given by

$$\begin{aligned} H_{nm} &= \text{Score}(X_1, \ldots X_n; Y_1, \ldots, Y_m) && (1) \\ &= \max_{i,j,\Delta} \sum_{k=0}^{\Delta-1} S(X_{i+k}, Y_{j+k}). && (2) \end{aligned}$$

The negative expectation value of S yields local alignments. Suppose it was positive, unrelated segments would obtain positive scores and the sum in (1) would tend to run over the entire sequence length. We restrict the analytical analysis to the gap free case (i.e the BLAST setup).

3 The Karlin–Altschul– (Dembo–Arratia–Waterman)– formula

The main result on the significance of alignment scores given the length of sequences is

$$P[H_{nm} > t] \approx 1 - \exp(-\gamma \, m \, n \, e^{-\frac{t}{\theta}}), \tag{3}$$

where γ and θ are parameters, that can be calculated analytically. θ is the positive solution of $E[\exp(S/\theta)] = 1$ and γ is hard to to write in TEX. If you want to know it anyway see the appendix in Karlin and Altschul (1990). This and similar results are discussed in the literature at several places. See Dembo and Karlin (1991b,a); Dembo *et al.* (1994); Karlin and Altschul (1990); Arratia *et al.* (1986, 1988) or for a comprehensive review Waterman (1995). This distributional result was preceded by the following strong law on the growth of alignment scores

$$\lim_{n\to\infty} \frac{H_{nn}}{\log(n^2)} = \theta \qquad \text{a.s..} \tag{4}$$

It is useful to reformulate equation (3) in terms of a regression formula:

$$H_{nm} \sim \alpha + \theta \, \log(n \, m) + \theta \, G. \tag{5}$$

Hence there is a linear dependence between the score and the log of the search space. The residuals obey a rescaled extreme value distribution i.e. $P[G < t] = \exp(-e^{-t})$. The slope of the regression line and the scale parameter of the residuals are identical. In a database search the length of the query remains the same in all comparisons, whereas the length of database entries varies. For the score H_i of a database entry of length n_i we get

$$H_i \sim \tilde{\alpha} + \theta \log(n_i) + \theta G. \tag{6}$$

To compensate for the bias resulting from different sequence length one can replace raw alignment scores by the p-values obtained from (3). These describe significance given sequence length. Alternatively one can measure similarity in terms of the residuals

$$A_i = H_i - \log(n_i)\,\theta.$$

Sorting for decreasing residuals and sorting for increasing p-values ends up in the same ranking.

The development of these results started in the mid 80's. Major contributions came from large deviation theory and from Poisson approximation. In the following we sketch some of them.

1. *Probabilities for high scoring fixed alignments:* **Cramér's Theorem.**
   A randomly picked alignment that scores high, can be interpreted as a large deviation from the law of large numbers. The classical theorem of Cramér (1938) provides a general theory for this kind of problems, and it is crucial for:

2. *Probabilities for rare blocks:* **Erdös–Rényi laws.**
   In alignment statistics we are not interested in the probability that a fixed alignment is high scoring, but in the probability that among all possible alignments between the two sequences at least one is high scoring. The results in Erdös and Rényi (1970) are a guideline for proving the strong law stated in (4). Applications of Erdös–Rényi Laws to sequence alignment are described in Arratia and Waterman (1989) a proof for the result stated in (4) is given in Dembo and Karlin (1991b).

3. *The distribution of scores:* **The Chen–Stein method or the Aldous clumping heuristic.**
   While Erdös–Rényi laws reveal the linear dependence of scores and the log of the search space, they do not yield the distribution of the residuals. The Chen–Stein method Chen (1975) yields a rigorous proof of (3), see Arratia *et al.* (1990b); Dembo and Karlin (1991a). A nice introduction to Poisson approximation via the Chen–Stein method including applications to

sequence comparison is given in Arratia *et al.* (1990a). Heuristic explanations of (3) can be obtained by using the Aldous clumping heuristic. This is done in Waterman and Vingron (1994) and in an even more elegant way in Ed's comments at the end of these notes.

4 Cramér's Theorem

Consider any randomly picked gap free alignment of two random sequences. What is the Probability that it scores high? This problem fits well into the context of Cramér's theorem. Cramér's theorem is discussed in many textbooks on probability. A modern and comprehensive introduction can be found in Dembo and Zeitouni (1992).

The setup: Consider a sequence of independent random variables $X_1, X_2, \ldots$ with common distribution μ on a finite set $\mathcal{M}$ and let S be a real valued function on $\mathcal{M}$ (In our case imagine the X_i to denote pairs of randomly aligned letters, $\mathcal{M}$ all pairs of letters and S the score function.)
By the Law of large numbers we have:

$$\frac{1}{n}\sum_{k=1}^{n} S(X_k) \longrightarrow E_\mu[S] =: m_0$$

For $m > m_0$ let

$$A_n(m) := \left\{ \frac{1}{n}\sum_{k=1}^{n} S(X_k) > m \right\}.$$

By the law of large numbers we clearly have

$$\lim_n P\left[A_n(m)\right] = 0.$$

Events of that type are called *large deviations*. Cramér's theorem provides probabilities for $A_n(m)$. We need some more notations in order to formulate the theorem.

For the reference distribution μ and the statistic S we define the exponential family of probability distributions given by

$$\mu_i^\lambda = \frac{e^{\lambda s_i}\,\mu_i}{\sum_i \mu_i\, e^{\lambda s_i}}\ ,$$

There is a one to one correspondence $m \leftrightarrow \mu^{\lambda(m)}$ of $m \in (\min S, \max S)$ and the probability distributions in $\{\mu^\lambda\}$, such that

$$\sum_i \mu_i^{\lambda(m)}\, S_i = m.$$

Note that $A_n(m)$ describes large deviation behavior with respect to μ whereas it is a normal event with respect to $\mu^{\lambda(m)}$. Let $I(m)$ be the relative entropy of $\mu^{\lambda(m)}$ with respect to μ i.e.

$$I(m) = H(\mu^{\lambda(m)} \mid \mu) = \sum_i \mu_i^{\lambda(m)} \; log \left(\frac{\mu_i}{\mu_i^{\lambda(m)}} \right).$$

We get:

Theorem: (Cramér 37)

$$\lim_{n\to\infty} \frac{1}{n} \log \left(P \left[A_n(m) \right] \right) = -I(m) \quad a.s. .$$

Probabilities of large deviations decrease exponentially: $P\left[A_n(m)\right] \approx \exp\left(-n\, I(m)\right)$. The exponential rate is determined by the entropy distance of the original model μ and a second model $\mu^{\lambda(m)}$ for which $A_n(m)$ describes normal (ergodic) behavior. A proof can be found in Dembo and Zeitouni (1992).

5 Rare blocks

In sequence comparison we chose among all possible alignments the one, that optimizes the score. Hence we are not interested in the probability that any fixed alignment is high scoring, but in the probability that among all possible alignments there is at least one that is high scoring. A one dimensional equivalent of this problem, is as follows: Consider an i.i.d. sequence of real valued random variables, what is the probability that there is a segment in this sequence with an average clearly above the expectation value of each variable. A mathematical framework for this problem is described in Erdös and Rényi (1970).

With the notations of the previous section, consider a long segment of the sequence $X_1, X_2, \ldots$ given by a starting–point k and an end–point $l \gg k$. By the law of large numbers we expect

$$\frac{1}{l-k} \sum_{i=k}^{l} S(X_i) \approx m_0.$$

For $m > m_0$ consider the events

$$C_{kl}(m) = \left\{ \frac{1}{l-k} \sum_{i=k}^{l} S(X_i) > m \right\}.$$

The C_{lk} are called rare blocks. For fixed (l,k) Cramér's theorem yields

$$P\left[C_{kl}(m)\right] \approx e^{-|l-k|\,I(m)} \quad . \tag{7}$$

Obviously long i.i.d. sequences contain some blocks of this kind, however their length is usually restricted. Erdös and Rényi (1970) describe the growth of rare blocks with the length of sequences:

Theorem: (Erdös Rényi 70)

$$\lim_{n\to\infty} \frac{R_n}{\log(n)} = \frac{1}{I(m)} \tag{8}$$

Consider the simple score function given by $S(a,a) = 1$ and $S(a,b) = -\infty$ if $a \neq b$ and suppose that gaps are not allowed. The optimal local alignment with respect to this score function is the longest common sub-word of both sequences. In this case (8) is identical to the strong law in (4). It also covers generalizations allowing for a given proportion of mismatches. For general scoring schemes random walk theory needs to be applied, see Dembo and Karlin (1991b). Their proof goes along the lines of the earlier results in Erdös and Rényi (1970), however their generalizations result in a technically much more difficult proof. The proof of Erdös and Rényi mainly involves Cramér's theorem and the Borel-Cantelli lemma. Before we start with the proof note the following duality. Let T_r be the waiting time for the first m-block of length r i.e.

$$T_r = \min\left\{ l \;:\; \frac{1}{r} \sum_{i=l-r}^{l} S(X_i) > m \right\}.$$

Clearly $\{R_n \geq r\} = \{T_r \leq n\}$. Instead of examining the length of rare blocks, we can as well study waiting times for rare blocks of a given length:

Lemma:

$$\lim_{n\to\infty} \frac{R_n}{\log(n)} = \lim_{r\to\infty} \frac{r}{\log(T_r)}.$$

Hence (8) translates to $\log(T_r)/r \longrightarrow I(m)$.
Proof of the lemma:

$$\frac{R_n}{\log(n)} \leq \frac{R_n}{\log(T_{R_n})},$$

because $T_{R_n} \leq n$. Hence

$$\limsup_n \frac{R_n}{\log(n)} \leq \limsup_r \frac{r}{\log(T_r)}.$$

On the other hand $T_{R_n+1} \geq n$ and hence

$$\frac{R_n}{\log(n)} \geq \frac{R_n}{\log(T_{R_n+1})} = \underbrace{\frac{R_n}{R_n+1}}_{\to 1} \frac{R_n+1}{\log(T_{R_n+1})}$$

Hence

$$\liminf_n \frac{R_n}{\log(n)} \geq \liminf_r \frac{r}{\log(T_r)}.$$

Now for the proof of the Erdös Renyi law:

$$\begin{aligned} \{T_r \leq n\} &= \bigcup_{k=0}^{n-r} \bigcup_{l=k+r}^{n} C_{kl}(m) \\ &\subset \bigcup_{k=0}^{n-1} \bigcup_{l=k+r}^{\infty} C_{kl}(m) \end{aligned}$$

With $p_\Delta = P[C_{lk}(m)]$, where $\Delta = |l-k|$ we get

$$P[T_r \leq n] \;\leq\; n \sum_{\Delta \geq r} p_\Delta.$$

Cramér's theorem yields a constant c such that

$$p_\Delta \leq c \, \exp(-\Delta \, (I(m) - \epsilon/2)), \tag{9}$$

for all $\Delta > 0$. Hence the sum in (9) is a geometric series and by including the limit into the constant we have

$$\begin{aligned} P[T_r \leq e^{r\,(I(m)-\epsilon)}] &\leq \widetilde{c}\, e^{r(I(m)-\epsilon)}\, e^{-r\,(I(m)-\epsilon/2)} \\ &= \widetilde{c}\, e^{-r\,\epsilon/2} \end{aligned}$$

Since $\sum_r P[T_r \leq \exp\,(r\,(I(m)-\epsilon))]$ is again a geometric series and hence finite, the Borel-Cantelli lemma yields that for large r we have $T_r > \exp\,(r\,(I(m)-\epsilon))$ and hence

$$\liminf_r \frac{T_r}{r} \geq I(m) \quad a.s.$$

So far we did not use the independence of the X_i. However for proving an upper bound we need independence. For $\alpha \in N$ and $n = \alpha\, r$ we dissect $X_1, X_2, \ldots X_n$ into successive non overlapping blocks of length r. Let $B_l = C_{(l-1)r,lr}$ be the event

that the l'th such block is a m-block. The B_l are independent events and their union is clearly included in $\{T_r \leq n\}$. Cramér's theorem yields

$$P[B_l] \geq c\, e^{-r(I(m)+\epsilon/2)},$$

for all r. Further $1 - x \leq \exp(-x)$ yields

$$P[T_r > n] \leq \Big(1 - P[B_1]\Big)^{\alpha} \leq e^{-\alpha P[B_1]}.$$

as above we get

$$\begin{aligned} P[T_r > e^{r\,(I(m)+\epsilon)}] &\leq \exp\left(-\frac{1}{r}\, e^{r\,(I(m)+\epsilon)}\, c\, e^{-r\,(I(m)+\epsilon/2)}\right) \\ &= \exp\left(-c\,\frac{1}{r}\, e^{r\,\epsilon/2}\right) \\ &\leq \exp\left(-c\, r\right) \qquad\qquad \text{for large } r \end{aligned}$$

The sum over the left hand probabilities is finite and the Borel-Cantelli lemma yields for large r

$$T_r \leq e^{r(I(m)+\epsilon)} \quad \text{a. s.}$$

and finally

$$\limsup_r \frac{\log(T_r)}{r} \leq I(m).$$

6 Distributional results

In order to prove that the residuals $H_{nm} - \log(nm)\,\theta$ obey an extreme value distribution, the Chen-Stein theorem is applied, see Arratia *et al.* (1990b); Dembo and Karlin (1991a). A very comprehensive discussion of these techniques in the context of sequence comparison is given in Waterman (1995). While this is again technically difficult, applying the Aldous clumping heuristic yields simple heuristic explanations for these results. Waterman and Vingron (1994) used this heuristic approach in order to extend the results to local alignment with gaps. And Ed sketched a very nice heuristic proof for the non gap case, that you can find in the appendix to these notes.

7 Gaps

Up to now we have discussed statistical results for non gapped alignments. However it has proved much more successful to do database searches using a real

Smith-Waterman algorithm. That means allowing for gaps. The statistics qualitatively depends on the gap penalties. Similar to the global vs local behavior for $E[S] > 0$ respectively $E[S] < 0$ there is also a phase transition between low and high gap costs. To illustrate this, suppose the cost for any mismatch were infinite. If we do not allow for gaps, the local alignment is the longest common sub-word of both sequences. We have shown, that it's length growth with the log of the search space. On the other hand, if we have free gaps the alignment is the longest common subsequence which is known to grow linearly. Very high gap costs yield similar behavior as infinite gap costs and on the other hand low gap cost alignments resemble the free gaps case. In fact, there is a phase transition line in the parameter space separating parameters (score matrix and gap penalty) that result in a logarithmic growth of the score from those that give rise to linear growth. A discussion of this result in the case of constant mismatch costs and linear gap penalties is given in Waterman (1995). Notice that these results all refer to the null hypothesis of uncorrelated i.i.d. sequences. Linear growth of alignment scores with the size of the search-space means in the context of database searches an explosion of noise. And database searches with gap costs in the linear regime do not work at all. Best results are obtained for relatively high gap costs in some distance to the phase transition line. The Karlin-Altschul formula is not proved for that case, however there is empirical evidence that it still yields reasonable results, at least after adjusting the parameters γ and θ, see Waterman and Vingron (1994).

References

Arratia, R., Goldstein, L. and Gordon, L. (1990a) Poisson approximation and the chen-stein method. *Statistical Science*, **5**, 403–434.

Arratia, R., Gordon, L. and Waterman, M. S. (1986) An extreme value theory for sequence matching. *Ann. Stat.*, **14**, 971–993.

Arratia, R., Gordon, L. and Waterman, M. S. (1990b) The Erdös–Rényi strong law in distribution, for coin tossing and sequence matching with a given proportion of mismatches. *Ann. Stat.*, **18**, 539–570.

Arratia, R., Morris, P. and Waterman, M. S. (1988) Stochastic scrabble: Large deviations for sequences with scores. *J. Appl. Prob.*, **25**, 106–119.

Arratia, R. and Waterman, M. S. (1989) The Erdös–Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab.*, **17**, 1152–1168.

Chen, L. (1975) Poisson approximation for dependent trials. *Ann. Probab.*, **3**, 534–545.

Cramér, H. (1938) Sur un nouveau théoreme-limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles, in Colloque consacré à la théorie des probabilités*, **736**, 5–23.

Dembo, A. and Karlin, S. (1991a) Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of markov variables. *Ann. Prob.*, **19**, 1756–1767.

Dembo, A. and Karlin, S. (1991b) Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables. *Ann. Prob.*, **19**, 1737–1755.

Dembo, A., Karlin, S. and Zeitouni, O. (1994) Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Prob.*, **22**, 2022–2039.

Dembo, A. and Zeitouni, O. (1992) *Large Deviation Techniques*. Jones and Bartlett, Boston, London.

Erdös, P. and Rényi, A. (1970) On a law of large numbers. *J. Anal. Math.*, **22**, 103–111.

Hwa, T. and Lässig, M. (1998) Optimal detection of sequence similarity by local alignment. In *Proc. of the second annual international conference on computational biology*, pages 109–116, New York NY, ACM.

Karlin, S. and Altschul, S. F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 2264–2268.

Waterman, M. S. (1995) *Introduction to Computational Biology*. Chapman and Hall, London.

Waterman, M. S. and Vingron, M. (1994) Sequence comparison significance and Poisson approximation. *Stat. Sci.*, **9**, 367.