Recapitulation and improvement of our sequence approaches published in years 2002 and 2004

Kubrycht J., Borecký J., Souček P., Ježek P., Růžička M., Sigler K.

Prague, 4 December, 2006

Contents

WP1. Introduction

Our system **ELEMS** (evaluation of sequence similarities using length equivalent measures) can be first of all employed in comparisons of motifs and site-restricted sequences of **short and very short lengths** (see also WP2.6). ELEMS represents a test for our hypothesis about the validity of column similarities of sequences. This means that we map several or multiple co-locating identities and similarities in different chains.

The sequence block is in our case evaluated as the sequence of columns. This enables us (in spite of some diversification) to use ELEMS in both multiple and double sequence alignments. ELEMS is as yet used as a post-alignment procedure. Nevertheless, some independent ELEMS-based alignment can be among others performed when using auto-projection of a single sequence set instead of alignment of two sequence sets described previously (Kubrycht et. al. 2002). The possibility to **grade principally different qualities of column** similarities belongs to the interesting advantages of ELEMS. We can distinguish minimal fuzzy-related similarities, widely used double sequence similarities and their multi-sequence analogues, cohesive similarities and “almost common” aa designated here as CCBE (important for PHI BLAST searches; WP2.3). “Sequence logos” represents the closest system

to ELEMS with respect to output information (Schneider and Stephens, 1990; Perez-Bercoff et al., 2006).

Besides a single column similarity we can indicate also **double column coincidences** of aa provided that partial (questionable aa related) and overall (dispersed template motif aa) limits are achieved (Kubrycht et al., 2002). The analysis of low-density three column aa coincidence is more questionable because of strong increase of corresponding limits with column number. Nevertheless, gap arrangement or simultaneous comparison of nucleotide sequences can represent an alternative solution in such cases.

In principle, ELEMS sorts column sequence similarities with the help of **aa density-related limits** independent of the regular range of lengths and heights representing cut-off limits of calculated length equivalents. We distinguish vertical length equivalents (WP2.1) and horizontal length equivalents (WP2.2). The former equivalents are related to a single column or column pair similarities or to the whole sequence block (mean length equivalent value, i.e. **"mean compressed" block height;** see also our ppt file). The latter ones detect overall block probability of nonrandom events and also enable us the necessary grading of sequence similarities between differently high and long sequence blocks as well as block selection according required to the validity (WP2.4 and WP2.5).

Random DNA approach was used in our first solution of ELEMS principles (version **ELEMS(RDA)**). This approach was based on a simplifying assumption of random occurrence of nucleotides determining corresponding aa probabilities. This model can be used mainly in the range of short chains, because it indicates non-random deviations from the most probable status **with respect to random mutation noise**. This is also why ELEMS(RDA) may serve as a parallel second or third possible method in more decisive cases of short peptides.

Both general and ELEMS(RDA)-specific formulas have been already published. General formulas appear to be suitable also for **knowledge-based marginal approaches** and perhaps also for joint probability evaluation (Kubrycht and Sigler, manuscript in preparation). On the other hand, ELEMS(RDA)-specific ones include, among others, otherwise underestimated relationships.

In this paper we not only summarize and specify essential rules of ELEMS published in the two preceding papers, but also show some more precise insights and reevaluations. In addition, we demonstrate here a comparison between results of ELEMS(RDA) and BLAST search for short sequences (WP3.2.2). We also compare the **limits independently derived by binomial ACS and Bayesian approaches** (WP3.2.3). Complicated enumeration of the

probability determining the absence of column similarity in ELEMS(RDA) is also shown as one of the obstacles of this otherwise lucid version of ELEMS.

WP2 Essence of general approaches determining vertical and horizontal length equivalents

WP2.1 Vertical length equivalents

Vertical length equivalents (**VLE**) concerning column height ($\mathbf{LE_A}$) and mean column height in sequence block ($\mathbf{LE_{TM}}$) can be calculated by the following formula:

$LE_A = \log(c_A) : \log(a_A)$, (WPF1)

$LE_{TM} = \log(b_{TM}) : \log(c_{TM})$, (WPF2)

where $\mathbf{c_A}$ is a (positional) column probability of aa (see also WP3.1.1) derived with the help of aa probability ($\mathbf{a_A}$) defined according to the used ELEMS version; $\mathbf{c_{TM}}$ is probability of template motif (**TM**) is the product of aa or alternative aa probabilities corresponding to the compressed structure of the required model twin block (see below); and $\mathbf{b_{TM}}$ is probability of aa similarities in given arrangement of a sequence block with or without gaps (Kubrycht et al., 2002; WP3.1.2).

In case of full aa identity in a column, LE_A always determines the number of compared chains (column height; Kubrycht et al., 2002). Less than full column identities determine non-integer heights of correlated model twin columns (**MTC**) containing exclusively aa identities or similarities. Such MTC heights can also indicate the stage of similarity, e.g. the height higher than three ($LE_A < 3$) suggests some additional structural linkages between aa in given positions (necessary column property in so-called **cohesive similarity**), the height of two is related to the frequently used **double sequence similarity** and the height one indicates **random chain** without linkages. Since the heights of actual columns containing full aa MSA similarity exhibit linear dependence on the height of the sequence block, we assume the same property in case of fully similar MTC of non-integer height. This means that each MTC column (related to MSA) of height $LE_A > 1.5$ thus contains aa (template motif aa, i.e. **tmaa**) more likely similar than dissimilar in accordance with the preceding remarks to heights two and one. Consequently, only aa achieving given tmaa level are evaluated in ELEMS, whereas residual non-random aa in the columns of LE_A = 1-1.5 (**questionable aa**) are usually specifically displayed as weakly reliable, being sometimes reevaluated in possible following

Table WPT1. Fuzzy-related intervals and the limiting values of ELEMS

	LE	tendency in interval	conditioned by
structures,properties	**fuzzy intervals**[a]		
questionable aa	$1.0 < LE_A \leq 1.5$	linear decrease	-
latent quasimotifs	$1.0 < LE_{TM} \leq 1.5$	linear decrease	$LE_A > 1.5$
stable template motifs	$1.5 < LE_{TM} \leq 3.0$	linear increase	$LE_A > 1.5$
similarity	$1.0 < LE_{A,TM} \leq 2.0$	linear increase	-
cohesivity	$2.0 < LE_{A,TM} \leq 3.0$	linear increase	presence of HS
	limited values[a]		
quasimotifs	$LE_A > 1.0$	-	only columns with $LE_A > 1.5$
	$LE_{TM} > 1.0$	-	are evaluated
	HLE > SL	-	
templates	$LE_A > 1.0$	-	only columns with $LE_A > 1.5$
	$LE_{TM} > 1.5$	-	are evaluated
	HLE > SL	-	
regular BCS (also tmaa)	$LE_A > 1.5$	-	-
acceptable blocks	see templates	-	-
DS sim.	HLE > SL ($LE_A = 2$)	-	-
high occurrence aa	$LE_A > 3.0$	-	-
newly designed structures and similarities[b]			
quasi-cohesive sim.	$HLE_{CoS} > SL$	-	$LE_{TM} > 2.0$
	$LE_A > 3.0$	-	-
H-stereotypes (**HS**)	$LE_A, LE_{CoS} \geq 3.0$	-	HS exist at least in
	$HLE_{CoS}, LE_C > SL$	-	three block segments
cohesive sim. (**CoS**)	$LE_A, LE_{TM} > 3.0$	-	at least half of seq.
	$HLE_{CoS} > SL$	-	form different HS
double sequence CoS	$HLE_{CDS} > 3 \times SL$	-	-
CCBE	$LE_A > SL + 2$	-	CoS block limit whithout given BCS
CoS blocks of the second order backward sim.	$LE_{TM} > SL + 2$	-	actual CCBE
	$HLE_{CoS} > SL$	-	see also CoS
	SIM(ALL) > 0	-	SHS define compared sequence block

[a]The tendencies and limits represent more likely a description of actual grading of aa/chain similarities, than a typical fuzzy system (Jura, 2003). Nevertheless, several ELEMS limits were proposed. For more selected choice of ELEMS similarities see sections WP2.1 and WP2.2.

[b]BCS - block column similarities; CDS - DS corresponding to CoS (see WP2.8); DS - double sequence similarity(-ies) including **motif** and **site restricted** sequence similarities, which both can be pre-selected by database search; SHS - sequences of high stability, i.e. domain sequences, standardized consensi, and reference sequences (Kubrycht et al., 2006); sim. - similarity. For details and explanation of other abbreviations see sections WP2.1-4, and WP4).

procedures (Kubrycht et al., 2004). In fact, the limit 1.5 represents a value derived by a simple defuzzification. Model twin block (**MTB**) more-likely similar than dissimilar can be restricted by $LE_{TM} > 1.5$ as well as MTC. Similarly to aa relationships in MTC definition, MTB is strictly constituted by neighbor aa. Probabilities of aa in mutation relationships are usually summed up and evaluated together (Kubrycht et al., 2002). On the other hand, probabilities of alternative blocks formed by alternative aa without mutation convertibility are usually summed up, whereas corresponding LE_{TM} calculation contains the product of alternative aa

probabilities raised to the fractions related to occurrences of alternative blocks. Besides this overall evaluation, we can also separately calculate b_{TM} variants for each sequence block, as it occurs in maximum likelihood comparisons.

Provided that reevaluation of sequences in block is necessary (e.g. in cases of a low number of columns, weak conditions for given alignment or when using chain matrix for pre-alignment; Kubrycht and Borecký, 1998) we check first of all the density of tmaa in selected chains (Kubrycht et al., 2002). The formula restricting a chain density of aa is similar to that enumerating a column length equivalent:

$$LE_C = \log(c_M) : \log(a_{GM}) > 1.5, \qquad \text{(WPF3)}$$

where c_M is chain probability of tmaa, $\mathbf{a_{GM}}$ is geometrical mean of tmaa probabilities and $\mathbf{LE_C}$ is a single chain-related length equivalent.

Chains of insufficient LE_C densities can be removed from pre-formed sequence block, which may cause a reduction of sequence blocks. Provided that this reduction influences the TM sequence, the following LE_C-related iterative refinement of chain density has to be started (Kubrycht, 2002). For more detailed description of ELEMS limits determining different types of similarities and accompanying fuzzy-related tendencies see Table WPT1.

WP2.2 Formulas determining horizontal length equivalent

Statistical evaluation of length equivalents usually restricts specific limits (**SL**) related to a_A or, more frequently, to a_{GM}. SL represents the length of chain determined by minimal significance. Since subtraction of random chain is necessary, the value LE_{TM} -1 represents mean density of non-random aa in the column, which includes topical gap evaluation. Consequently, we can calculate the length of non-random twin chain substituting sequence block. This length is named horizontal length equivalent (**HLE**):

$$HLE = (LE_{TM} - 1) \times k = \log (b_{TM} / c_{TM}) : \log(a_{GM}) > SL, \qquad \text{(WPF4)}$$

$$INF(HLE) = SL, \qquad \text{(WPF5)}$$

Where **INF** denotes unachievable inferior possible value of HLE. HLE of sequence block have to be thus higher than is SL length of minimal single independent chain of the same a_{GM}.

WP2.3 Two possible levels of more dense sequence similarities in more detail

Cohesive similarity (**CoS**) represents a strict type of similarity related to a higher level than TM-derived one. Among others all cohesive model twin blocks (**MTB**) and model twin columns (**MTC**; for both model items see WP2.1) keep TM similarity even when losing one

of their model chains or aa, respectively (i.e. each limiting double sequence similarity of MTB or MTC has to be confirmed by more than one cohesive random chain or one cohesive column aa in case of CoS, respectively). CoS is defined by the formula similar to but more strict than formula WPF4:

$HLE_{CoS} = (LE_{CoS} - 2) \times k = \log (b_{TM}/c_{TM}^2) : \log(a_{GM}) > SL,$ (WPF6)

CCBE (column similarities of cohesive block extent) related aa represent milder variants of common aa. Their limit SL + 2 follows from formula WPF4 similarly to relationships derived in WP2.8. Like common aa, CCBE-related aa are candidates for pattern (alternative pattern) aa used in PHI BLAST searches. For additional restrictions of CoS and CCBE-related similarities see Table WPT1.

WP2.4 Calculation of specific limits

Our statistical evaluation uses two distinct approaches to determine SL, i.e. binomial evaluation. Binomial evaluation determines the significance of aa probability (**p**) using a formula for binomial sample size (**s**) (Lepš, 1996):

$s = p \times (1-p)/w^2,$ (WPF7)

where w is level of significance (P < w).

In accordance with our approximation of chain slices (**ACS**; Kubrycht et al., 2004) each aa in compared sequence represents also the start of the following chain of length M, (i.e. we can also consider $p = (a_{GM})^M$) and this probability can also determine chains of non-integral lengths (slices) such as HLE. Consequently we may assume the formula for one such slice (s = 1):

$w = ((a_{GM})^{HLE} \times (1- (a_{GM})^{HLE}))^{1/2}$. (WPF8)

When using usual statistical limit w = 0.05, and formula WPF5, the formula is converted to equation (for a_{GM} and LE_{TM} see also WP2.1):

$2.5 \times 10^{-3} = (a_{GM})^{SL} \times (1-(a_{GM})^{SL})$. (WPF9)

Bayesian evaluation is less specific with respect to our problem but its evidence follows directly from widely-known formulas (Komenda, 1997; Zimmer, 2004). To evaluate the relationship of null (NULL) and biased (BIA) model we assume arrive from the formula:

$P(NULL| p) = (P(p| NULL) \times P(NULL)) : P(p),$ (WPF10)

where P(p) = P(p| NULL) x P(NULL) + P(p| BIA) x P(BIA). This means that the limiting value w = 0.05 (when P(NULL) = 1 - w and P(BIA) = P(NULL| p) = w) determines a more specific equation determining SL:

$0.05 = 0.95 \times (a_{GM})^{SL} : (0.95 \times (a_{GM})^{SL} + 0.05 \times (1 - (a_{GM})^{SL})$, (WPF11)

when considering $P(p| NULL) = (a_{GM})^{SL}$ and $P(p|BIA) = 1 - (a_{GM})^{SL}$.

Bayesian evaluation yields very related SL values. For details see WP3.2.3 and Table WPT2.

WP 2.5 Unification of specific limits and block-related horizontal length equivalent.

To compare any block sequence similarities with calculated HLE values we define significant relative block similarity (**RBS**):

$RBS = HLE/SL > 1$ (more precisely $HLE(a_{GM};w)$ and $SL(a_{GM};w)$). (WPF12)

Maximum positive RBS value is decisive when we select "the most stable" sequence TM following from evaluated sequence block similarly to maximum likelihood evaluation (Tateno et al., 1994). Though "the most stable" TM can be important in our case, it does not always represent the sequence of the most specific structure with respect to the solved problem. Hence we can expect hybrid similarity of considered ancestor to both reference and selected block sequences (in agreement with hybrid template selection described in Kubrycht et al., 2006).

In addition to TM similarities of $RBS > 1.0$ and $LE_{TM} > 1.5$, the same RBS limit restricts also the lower dense but still valid sequence similarities of $LE_{TM} > 1.0$. These similarities yield quasimotif consensus sequences (see also WPT1 or Kubrycht et al., 2002) and include the majority of usual ELEMS-related double sequence blocks. The latter possibility follows from the dissipative behavior of all-or-none law based double sequence comparison (Kubrycht et al. 2002), which appears to be a good reason for the usage of HLE in the quasimotif context in such cases.

WP2.6 Problems with usage of ELEMS in case of more extended sequence blocks

Only the limited extents of columns and chains can be evaluated by the first ELEMS variant (ELEMS(RDA)). Such restriction can be considered based on decreasing percentage in model similarities (Table WPT2) and model columns (Kubrycht et al., 2002), and existence of sequence repetitions first of all on levels repeats and domains.

. Consequently, recent ELEMS can be used in the range of chain lengths and columns heights from two (for some restriction see WP2.7) to twenty in exact cases, to thirty in feedback and heuristic approaches but hardly in the cases of the extents of more than forty. The solution for longer chains can consist in changing the employed statistical distribution or in using an island-like strategy. The latter alternative means that dense regular similarities of

possible lower lengths (e.g. segments of at least 50% of aa similarity limited by SL* described in WP2.8) can be integrated to longer similarities as islands (Olsen et al., 1999; Poleksic et al., 2005). This alternative is among others also used in some BLAST programs (Altschul et al., 2001). The other solution perhaps follows from the following consideration.

Since domain repeats and domains exist, we can indicate an increasing number of background-related similar structures with increasing length. Such effect can be perhaps compensated by addition of cohesive floors in model twin block (**MTB**). This possibility would also explain parallelism between the two types of ELEMS(RDA) similarities and both BLAST searches described in Table WPT2. Corresponding hypothetical MTB changes can then be expressed by the hypothetical formula:

$$HLE = (LE_{TM} - (1 + r(N;a_{GM}) + d(N;a_{GM}))) \times n, \qquad \text{(WPF13)}$$

where $r(N;a_{GM})$ and $d(N;a_{GM})$ are functions related to repeat and domain similarities, respectively, N is length and a_{GM} is geometrical mean of aa probabilities. Values $r(N;a_{GM})$ and $d(N;a_{GM})$ perhaps depend on entropy evaluation similarly to constant H considered in the case of BLAST-related theoretical analysis (Altschul, 1991).

WP2.7 Blocks of critical heights with respect to evaluation of sequence similarities

Extremely low percentage of model sequence similarities (Kubrycht et al., 2002) and inferior values of c_A with respect to length equivalent limits can be observed in sequence blocks of height four. Consequently, we do not recommended to use the blocks of such extent and also too random arrangements of three sequences in procedures generating templates and template motifs. Similarly, double sequence blocks (**DS**; i.e. double sequence similarities) cannot be separately employed to derive given structures. Yet, DS of ELEMS appear to be still useful for the detection of similarities between two motifs or between motif and pre-selected database sequences (**motif similarities**). The closest sequence blocks to the given critical group, i.e. the blocks of heights five and six, exhibit also some losses of similarities in comparison with higher blocks and possibly also accompanying fluctuation of similarities (Kubrycht et al., 2002). Therefore we recommend to use such blocks only in necessary cases of well-defined pre-selected representative chains. The probability of absence of any aa similarity in a column without gaps ($\mathbf{Q_c}$) value has to be one rather than lower in such cases as well as in original strict approach $Q_c = 1$ (Kubrycht et al., 2002) or zero value in case of column score.

WP2.8 Consequences of inferior HLE relationship to SL

Since SL = INF(HLE) (see WP2.2), three types of useful limits can be derived when employing inferior relationships between HLE and SL. Limits for similarities with uncertain edges (**SL***) are in accordance with definition of TSB2 block (Kubrycht et al., 2002). The approach resembles Laplace Estimator correction used in Bayesian evaluation (Zhang, 2005). In principle we add to both TM edges half of column with minimal similarity (resulting in **HLE***) related to a_{GM} value:

$SL^*/SL = INF(HLE^*/HLE) = INF\ ((LE_{TM} - 1) \times k + 2 \times 1/2 \times (LE_{TM} - 1))/\ ((LE_{TM} -1) \times k)) = (k/2 + 2 \times 1/2 \times 1/2) : (1/2\ k) = (k +1)/k.$ (WPF14)

SL* can be used instead of SL in the corresponding preceding formulas. Since for any authentic SL holds SL > 1 (i.e. length of one aa), **at least three aa form double sequence similarities with uncertain edges**. Interestingly, this directly derived limit is independent of the use of special ELEMS version.

To enumerate specific limits related to cohesive double sequence similarity (**SL_{CDS}**), we considered the fact that LE_{TM} determining SL similarity (**LE_{SL}**) can be seen as a value following from double sequence MTB including two identical chains of SL lengths (i.e. $LE_{SL} = 2$). On the other hand minimum MTB related to cohesive similarity (**CoS**) contained infinitesimally more then three such chains ($LE_{CoS} = 3$). This meant that:

$SL_{CoS}/SL = INF((LE_{CoS} - 1) \times k)/\ INF((LE_{SL} - 1) \times k) = (3-1)/(2 -1) = 2,$ (WPF15)

where k is the number of similar columns in CoS and corresponding double sequence similarity of SL. In the second step we assumed independent SL limit for usual double sequence similarities, i.e.:

$HLE_{CDS} > SL_{CDS} = SL_{CoS} + SL = 3 \times SL.$ (WPF16)

A similar procedure determined the minimal number of columns in case of extremely low but sufficient column densities of sequence blocks (**k_{minLD}**). This value following from formulas WPF17 and WPF18 represented a good illustrative consequence of SL limits.

$SL/SL = INF((LE_{TM} - 1) \times k)/SL = ((1.5 -1) \times k)/SL = k/(2 \times SL)\ (= 1),$ (WPF17)

$k_{minLD} = ceil(k) = ceil(2 \times SL).$ (WPF18)

Since the number of actual columns has to be an integer we used C program function ceil(x) rounding up the values to the closest integer higher one in formula WPF18. For more concrete processing of all given three limits in version ELEMS(RDA) see WP3.2.1-2.

WP3 ELEMS(RDA) - the first ELEMS version

WP3.1 Main procedures and values

WP3.1.1 Random DNA-derived approach to aa probabilities. The first simplified solution of ELEMS, i.e. ELEMS(RDA) is based on a random DNA approach (**RDA**) described before (Kubrycht et al., 2002). In agreement with this approach random DNA sequence is assumed, i.e. the number of different triplets encoding each aa (**q**) enables us to calculate corresponding probability (**p**) of aa occurrence in each site (p = q/61; sixty one are used, because this is the number of all coding triplets). ELEMS(RDA) includes also evaluation of single triplet mutation aa alternatives as alternative elementary events (see below or Kubrycht et al., 2002). We distinguish positional (block) column (**PCS**) and (topical) block column (**BCS**) similarities in ELEMS(RDA) (Kubrycht et al. 2002). The probability of each aa in positional block column ($\mathbf{c_A}$), which is related to potential PCS, can be calculated using the formula:

$$c_A = HCs \times a_A^{\,s} \times (1- a_A)^{H-s} \qquad \text{(WPF19)}$$

H is the number of chains penetrating given column (in case of non-gapped similarities) or the number of chains forming a sequence block (height in case of sequence similarities with gaps), C denotes the number of combinations formed by s-positions at H sites, and s is a number of similar or identical aa in the column, a_A is aa probability q/61 mentioned above.

On condition that c_A determines LE_A higher than 1.5 (for details see formulas WPF1 and WPF19), the number of aa identities/positivities can be considered to be sufficient for PCS. In accordance with formula WPF19 PCS is independent of random insertion/deletion events, i.e. it does not depend on variously interpreted gaps. This relationship is in accordance with the fact that gaps are first of all products of block alignment. From another point of view we can also say that the loss of any block position (gap) does not mean the loss of chain (diminishing of height) at the same site. In spite of this, in the first selection step of column similarities the disregarded gaps are not removed from our evaluation but they only become to be reckon during the following second step, when probability related to block similarity is determined. During this second step gap and also BCS related c_{ABj} values are used instead of c_{Ai} ones in a fraction of columns exhibiting PCS to enumerate b_{TM} (see bellow).

WP3.1.2 Calculation of sequence blocks with gaps using ELEMS(RDA). The probability of columns ($\mathbf{Q_c(g_i)}$) without positional column similarity of aa (**PCS**; see also WP3.1.1) can be expressed by the formula:

$$Q_c(g_i) = G_i \times Q_c = HC(H-g_i) \times Q_c, \qquad \text{(WPF20)}$$

where G_i is gap member in the formula, H is column height as in WP3.1.1, g_i is the number of gaps in a given column, and C denotes the calculated number of combinations yielded by occurrence of $(H\text{-}g_i)$ aa positions at H sites and Q_c is a probability of absence any similarity in a column without gaps.

Similarly, the block column probability (identical with positional column similarity in case of non-gapped similarity (Kubrycht et al., 2002)) is determined by the formula:

$$c_{ABj} = c_{AB}(g_j) = HC(H\text{-}g_j) \times c_A(H\text{-}g_j) \qquad \text{(WPF21)}$$

where $c_A(H\text{-}g_i)$ is a positional column probability in block of height $H\text{-}g_i$ (for other details see WPF2, and WPF5).

Since G_i is separable from Q_c in formula WPF20 and c_{ABj} represents only a special (low weight) construct non-related to the common background of column similarities (see also WP3.3.3), the following formula for gapped block similarity is proposed:

$$b_{TM} = nCk \times Q_c^{n-k} \times \prod_{(j = 1 \text{ to } k)} (c_{ABj}) \times \prod_{(i = 1 \text{ to } n-k)} G_i = nCk \times \prod_{(j = 1 \text{ to } k)} (c_{ABj}) \times \prod_{(i = 1 \text{ to } n-k)} (Q_c(G_i)) \qquad \text{(WPF22)}$$

where n, k are numbers of all columns or columns including aa similarity, respectively, and G_i value is determined by WPF5. The values $G_i = 1$ and $c_{ABj} = c_{Aj}$ can be used in each case of non-gapped column. WPF7 is the explicit form of the formula established and used in one of our previous papers (Kubrycht et al. 2002).

WP3.1.3. Absence of aa similarity in a column represented by Q_c value. Topical column probabilities of aa similarities are not usually identical with limiting values, but are substantially lower and make jagged ("limit-overflow" related) oscillations near the limits. These oscillations are synchronous in the minimal sequence blocks (see also web page subsection WP2.7) and then become more asynchronous with respect to differently probable aa (for details see Kubrycht et al., 2002). Consequently, first of all we obtain a substantial increase of probability values related to aa similarities in model sequence block columns to the height of seven and then combined oscillation around almost the same mean value. The blocks of different height (**H**) therefore exhibit also oscillating values **$Q_c(H)$**:

$$Q_c(H) = 1 - P_c(H) = 1 - (\sum_{(i = 1 \text{ to } 5)} (\sum_{(j = 1 \text{ to } LAST1(j))} e_{ij}\, c_{Ai}^*(H) + \sum_{(j = 1 \text{ to } LAST2(j))} e_{Mij}\, u_{ij}\, d_{Ai}^*(H)), \qquad \text{(WPF23)}$$

where P_c is the probability of any aa similarity in the sequence column; $c_{Ai}^*(H)$, $d_{Ai}^*(H)$ are probabilities of aa column similarity related to each individual aa, i.e. these values concern separately evaluated column identities or additional single triplet mutation-derived similarities (**AMS**), respectively; e_{ij}, e_{Mij} are the proportions of j-th topical compartment of aa

convertibility in i-th aa level (five levels can be derived by ELEMS(RDA)) of probability, which correspond to evaluated aa identities or single triplet mutation-associated codon units (**MACU**), respectively; u_{ij} is i- and j-restricted probability of three alternative mutation relationships (values zero, 1/3, 2/3 and one are in accordance with ELEMS(RDA) assumptions); LAST1(i) and LAST2(i) are the number of aa or MACU in the i-th level.

Since we used the LE_A evaluation, the determination of geometrical mean of $Q_c(H)$ values appears to be a suitable procedure to estimate Q_c described in formula WPF22 and common for the selected oscillating $Q_c(H)$ values (see above). In agreement with usual statistical procedures, prevention of false positivities and restriction of random coincidences of oscillations, we used the following formula to determine such Q_c:

$$Q_c = MAX\ (\Pi_{(H = I\ to\ I + 2\ x\ MIN(U) - 1)} Q_c(H)/MIN(Q_c(H))^{1/(2\ x\ MIN(U) - 1)}), \qquad (WPF24)$$

where MAX, MIN are selected maximum and minimum values; U are the oscillation periods of the limiting column similarities related to two aa groups of the best contribution to $Q_c(H)$ values. MAX(U) and MIN(U) are nine and six, respectively (Kubrycht et al., 2002), whereas the character I denotes the values from seven to fifteen (i.e. 7 + MAX(U) - 1) in our case.

Besides the Q_c value enumerated by formula WPF24, additional reevaluation is necessary to estimate $\mathbf{Q_{c2}}$ value related to a precisely (non-asymptotically) limited sequence block of height two (double sequence similarity), and the Q_c values of asymptotically limited critical blocks of low heights outside the oscillation range described above (here we usually postulate $Q_{cCRIT} = 1$). For details related to preceding calculations or restrictions in given subsection see WP2.6-7, and Table WPT3). When disregarding the evaluation of AMS, only column absence of aa identities is evaluated and thus we obtain the additional values $\mathbf{Q_{cI}}$ and $\mathbf{Q_{c2I}}$ (= $Q_{cI}(2)$).

WP3.2 Examples of some interesting quantitative relationships

WP3.2.1 Similarities in sequence blocks of minimum LE_{TM}. In accordance with formula WPF18 a two times higher number of column similarities than determined by the SL value is necessary in sequence blocks with minimum LE_{TM} value (**minimum blocks**). The maximum q_{GM} value 3.053 restricts minimum number of four columns with aa similarity in minimum blocks (see also Table WPT2), whereas six column similarities are necessary for any minimum block currently evaluated by ELEMS(RDA). The limits corresponding to uneven numbers of column similarities then follow from half-related SL values. The minimum blocks with three and five column similarities are restricted by values $q_{GM} < 1.125$ (only identities of

tryptophan or methionine) and $q_{GM} < 5.558$ (almost all corresponding chain identities and majority of chain similarities derived by single triplet mutation), respectively.

WP3.2.2 Comparison of model limiting double sequence similarities determined by ELEMS with those following from BLAST searches. To illustrate the extension of chain similarities done by ELEMS(RDA) we calculated minimal lengths of non-gapped double sequence similarities (Table WPT2). The counterpart BLAST-related model examples were

Table 2. Model non-gapped double sequence similarities

Calculated limiting ELEMS similarities[a,b]												
q_{GM}/model	SL[c]	similarities								cohesive similarities		
q values	ACS	Bayesian	determined edges			uncertain edges						
			1/2	1/3	1/4	1/2	1/3	1/4		1/2	1/3	1/4
1	1.457	1.434	4	9	12	6	9	16		12	24	36
2	1.753	1.724	6	9	16	8	12	20		16	33	52
3	1.989	1.956	**6**	**12**	24	**8**	**15**	28		20	42	76
4	2.199	2.163	**8**	**18**	32	**10**	**21**	36		24	54	108
5	2.395	2.356	**10**	**21**	48	**12**	**24**	52		28	72	172
6	2.583	2.541	10	27	84	12	30	92		34	99	340

Examples of limiting modified similarities generated by BLAST program SNEM[a,b]									
strategies:		STR1		STR2		STR3		STR4	
num seq STR:		227541		19369		**1975**		78	
lengths values		int	non-int	int	non-int	int	non-int	int	non-int
6	%	83.3	74.8	66.7	66.2	66.7	**52.9**	n.f.	-
	q_{GM}	4.704	4.756	4.899	4.892	4.899	4.648	-	-
10	%	60.0	57.8	**50.0**	**49.1**	50.0	48.3	50.0	48.4
	q_{GM}	3.813	3.806	3.776	3.746	4.095	4.039	4.338	4.291
15	%	53.3	53.2	46.7	46.1	46.7	**42.6**	**40.0**	**37.3**
	q_{GM}	3.722	3.722	3.684	3.680	3.684	3.519	3.397	3.257

[a]A close relationship between ELEMS and SNEM BLAST limited similarities can be observed in the range of restricted but sufficient numbers of compared sequences and lengths (selected values are in bold). The majority of geometrical mean q values (**q_{GM}**) related to usual peptide sequences occur in interval from three to five. Displayed SL-related values (**Z**) are slightly higher than SL, i.e. $Z > SL \geq Z - 0.001$. For Standard Protein BLAST relationship to displayed cohesive similarities see 3.2.3. For additional details see 3. 1.2.

[b]1/2, 1/3, 1/4 - minimal lengths of sequences achieving accurate one half, third and quarter similarities are displayed, respectively; int, non-int - integer and non-integer numbers of similar aa were considered in our calculation; n. f. - not found; num seq STR - number of sequences restricted by given strategy; SNEM BLAST - BLAST search for short nearly exact matches; STR1-4 and STRIg - search strategies, i.e. STR1: Metazoa[ORGN] NOT Vertebrata[ORGN], STR2: Caenorhabditis briggsae[ORGN], STRIg: immunoglobulin[TW] OR Ig[TW], STR3: STR1 AND STRIg, STR4: STR2 AND STRIg.

[c]For all half-related and integer values of q_{GM}/model q values in the range from one to six, linear regression analysis of slightly more strict ACS-derived SL(q) determines r = 0.996788, a = 0.220000, b = 1.297732, MSe = 9.5452×10^{-4}.

generated with the help of BLAST searches in accordance with WP3.4.1 (Table WPT2). N-terminal hexapeptide, decapeptide, pentadecapeptide (maximum chain length in SNEM BLAST search) and the whole chain of conserved variable Ig domain cd00099 were compared with four different sequence sets. The data derived from the searches with given short peptide segments suggest that the sequence set derived by the third strategy in Table WPT2 (**STR3**; set with 1 975 sequences) appears to be in agreement with ELEMS limits. This means that about sixty times more extended set of longer sequences than usual ELEMS segments can be still processed by our limits. The searches based on STR3 revealed also at least some

dipeptides among minimal similarities (e.g. SC-SC), whereas only minimal tetrapeptide similarities were found by the closest but more strict STR2. The Standard Protein BLAST (**SPrB**) search with whole Ig domain query was performed only on a pre-selected STR3-related sequence set. This search resulted in two similarities without gaps closely related to cohesive double sequence similarities described in Table WPT2. The chains of lengths forty and thirty two aa achieved 42.5% and 46.9% SPrB similarities, of recalculated Expect values 2.3 and 2.5 (originally 6.3 and 8.3) and q_{GM} values 4.420 and 4.363, respectively. This means that a two-level parallelism exists between usual and cohesive ELEMS similarities and SNEM BLAST and SPrB in the different ranges of sequence lengths, respectively (see also WP2.6). In summary presented Table WPT2 related to BLAST exhibits agreement in range of chain lengths selected by bold.

WP3.2.3 Both ACS and Bayesian statistics determine almost identical specific limits in a broad range of aa probabilities. Since aa of maximum q = 6 are encoded by degenerated triplets, the maximum extent of their mutation compartment (see below) is done by three codon sets of q = 6. This means that the resulting compartment exhibits q = 18, which corresponds to maximum probability following from mutation relationships. In accordance with usual resolution requirement following from theory of information (Baier, 1972) we used q = 0.5 as limit from minimum probability.

All relative differences between SL values derived by ACS and Bayesian approaches calculated in the range of q = 0.5 – 18 selected above (see also Table WPT2) were less than 2%, This result suggested a good agreement between the two enumerations of SL in a sufficiently broad range of values. Though we employed ELEMS(RDA) approach in this subsection, presented evaluation exceeds frame of this approach being also usable in other current marginal approaches.

WP3.3 Details necessary for Q_c enumeration and determined Q_c values

WP3.3.1 Compartmentation of alternative aa subsets. If we assume a model restriction of aa convertibilities in case of evolution relationships, then a column occurrence of each aa represents an alternative event only to a part of occurrences of other aa forming thus a model compartment subset (**MCS**). Since aa of the complementary subset are not present in each column including MCS, they have to be present in a different column. This means that each MCS as well as its complementary set can be expected in any block column only with some probability less than one. The value of such probability corresponds to the fraction of aa forming MCS in the simplest cases. For instance, ELEMS(RDA) enables us to enumerate this fraction as a relative frequency of codons in a genetic code table. In more detail, each codon aa fraction includes codons of model source (modified) aa and codons

encoding aa accessible from this source aa via single or double (more than 95% are usually degenerated) triplet mutations, and via AT or GC transversions. Each such codon fractions represents then e_{ij} values necessary for calculation of formula WPF23.

*WP3.3.2 Additional single triplet mutation-derived similarities (**AMS**).* Except for the more advanced analysis of mutation clusters, the column occurrence of AMS is conditioned by the presence of at least two aa subsets whose codons are at least partially reciprocally convertible via a single triplet mutation and achieve separately questionable but not higher template motif aa level. Convertible codons (all aa codons or part of them) of each such aa subset then determine the frequency of AMS and represent mutation-associated codon units (**MACU**) (for additional comments see Kubrycht et al., 2002).

We have to recall, that probabilities of questionable similarities are lower then aa probabilities ($LE_A \geq 1$), whose arithmetical and geometrical mean values are equal and lower than 0.05 when using ELEMS(RDA), respectively. In addition, AMS-related aa pairs represent only a minor contribution to aa probabilities in our calculation of Q_c. In accordance with these facts, the ternary and higher multiple occurrences of AMS are not included to our background evaluation and construction of corresponding aa arrangement (the resulting Q_c is only slightly more strict with respect to false positivities).

The difference between the questionable and the TM levels of column aa similarity usually permits occurrence of unambiguous frequency of aa in the column of given height. In some cases this aa similarity level does not even exist. Since AMS form alternative elementary events (decreasing thus LE_A), their codon and column numbers are summed up in the corresponding block evaluation. Because of the usual double-frequency of AMS-related questionable aa pairs, the calculated column probability of AMS is always passed through the current restriction length equivalent limit (for relationship between frequencies and column heights see Kubrycht et al., 2002).

WP3.3.3 Model arrangement of AMS. Several principles were employed to construct a simple randomized model arrangement approximating discrete distribution of AMS probabilities. In accordance with the conclusions of preceding paragraph (WP3.3.2), only aa similarities between two aa sets were included in our arrangement. Since each codon can be mutated in three nucleotides, three alternative events of one third of probability can be related to each MACU (see also non-integer values in Table WPT3). However, only mutations in the first two nucleotides of codons usually result in aa change and AMS. The third nucleotide participates only in AMS of Trp and Met (zero value of u_{ij} in formula WPF23 is related to other aa in such case). This is one of neutral mutation effects mentioned also in Table WPT3.

The usage of MACU (see WP3.3.2) in our model arrangement comprises all existing single triplet mutation relationships between aa including the similarities between non-reciprocally convertible aa. More special calculation of non-reciprocal aa similarities is thus more likely the matter of topical evaluation of each aa sequence block column and our additional knowledge (see also Kubrycht et al., 2002), rather than the matter of the necessary background analysis. The model arrangements concerning both aa identities and AMS are displayed in Table WPT3 present in the following page.

Table WPT3. Important values of our model arrangements[a]

aa identities					
q values	1	2	3	4	6
number of aa	2	9	1	5	3
61 x $\sum_{(j = 1 \text{ to } LAST(i)} e_{ij}$	14	118	19	120	92
AMS					
original q values	1	2	3	4	6
q_{MUT}	2	4	6	8	-
Last(i) x w_{ij}	8/3	32/3	4/3	8/3	no MACU
61 x $\sum_{(j = 1 \text{ to } LAST(i)} e_{Mij}w_{ij}$	27	205	104/3	254/3	-

[a]Four additional factors influence the frequencies related to AMS arrangement in contrast to aa identities. Presence of **neutral mutation** and **STOP codons** diminishes a MACU number with respect to the corresponding number of aa. On the other hand, the **division of codons determining hexatriplet aa** to lower units compensates such MACU losses. In addition, a **larger extent of united (overlapping) aa compartments** formed by each aa pair composing AMS increases proportion of AMS with respect to the displayed arrangement of aa identities.

WP3.3.4 The values c_{Ai}(H) and d_{Ai}*(H) of formula WPF23 in more detail.* The very important c_{Ai}* values denote the column probabilities of aa identities restricted by the LE_A limits. Since each column can include only unique similarity of given aa, we may propose the formula:

$$c_{A\,i}^*(H) = (\sum_{(k = MIN(H) \text{ to } H)} c_{Aik}^2(H))/(\sum_{(k = MIN(H) \text{ to } H)} c_{Aik}(H)), \quad \text{(WPF25)}$$

where k is the number of identical aa in evaluated block column and $c_{Aik} = c_{Aik}(H; a_k; k)$

The additional value d_{Ai}*(H) represents c_{Ai}*(H) value related to AMS. In principle this value can be calculated like c_{Ai}*(H). However, reduced formula is sufficient in most cases:

$$d_{A\,i}^*(H) = d_{Ai}(H; 2a; k = L + M), \quad \text{(WPF26)}$$

where L and M are the aa numbers of aa in both subsets of questionable aa (usually L = M). Alternative still frequent absence of questionable level of similarity then determines $d_{Ai}(H)^* = 0$.

WP3.3.5 Stability, accuracy and oscillation effects in our calculation of Q_c values. Since the increase of Q_c diminishes RBS of compared similarities, such increased values provide a higher stability of calculation with respect to false positivities. Since accurate calculations are simultaneously required we usually have to restrict the effort to stabilize our evaluation of sequence comparisons only to the narrow interval surrounding mean or geometrical mean values. The statistical characteristics such as SD, SSD and confidence intervals can help us in this restriction. The period extent has to be further included in our consideration when analyzing oscillating events. The sample values of more than one period origin have to be then processed when the sum of several differently oscillating events is enumerated. Such restriction also concerns our formula WPF24.

WP3.3.6 Problems with regression analysis of oscillating $Q_c(H)$ values. The possible usage of regression analysis is complicated by at least three factors: i) besides the confidence interval relationships described in WP3.3.8, the model one period intervals located in the counterpart edges of calculated height range (representing height extents 7-12 and 21-26) yielded the differences (the values 6.5×10^{-3}, 9.9×10^{-3}, 1.1×10^{-3}, 1.5×10^{-3})

comparable with those following from logarithmic SD and SSD values; ii) the result of successful regression analysis has to be always stabilized with respect to random effects of oscillation, height range selection and asymptotic nature of the limits, iii) absence of the starting point to include three preceding combined random effects to a single model of regression analysis.

WP3.3.7 Q_{c2} value necessary for calculation of double sequence similarity. Since double sequence similarity (**DS**) achieves unambiguously $LE_A = 2$, the similarities are not asymptotic but deterministic in such case. This means that the precise $Q_c(2)$ value can be calculated. Nevertheless, to keep a model ELEMS(RDA)-related strictness of DS with respect to AMS occurrence (loss of this relationship is caused by enlarged deterministic LE_A related to aa identities of DS), we have to multiply each d_{Ai}* member of the sum in formula WPF23 by the corresponding proportion of probabilities. The proportion $a_{Ai}^{0.5}$ is topical for DS. In addition, the disjunction of AMS and aa identities has to be also included in the given enumeration. In the end, we thus gain adapted Q_{c2} value different from $Q_c(2)$, whereas $Q_{c2I} = Q_{cI}(2)$.

WP3.3.8 Determination of more precise Q_c and Q_{cI} values necessary for ELEMS(RDA) calculation. Using the approach described in WP3.1.3, we determined $Q_c = 0.9569235$, $Q_{cI} = 0.9622553$ in the enumerated range of heights (both selected maxima of formula WPF24 occurred also in the interval of heights from seven to twenty proposed for exact approaches in section WP2.6), and also more special $Q_{c2} = 0.9645645$ (this value is used in Table WPT2) and $Q_{c2I} = 0.9740551$. The statistical evaluation of geometrical means concerning asymptotic Q_c and Q_{cI} revealed the values 1000 x log Q_c = - 19.123 $\pm$ 3.898 and 1000 x log Q_{cI} = - 16.710 $\pm$ 3.366. Both resulting logarithmic SSD (selected standard deviations, i.e. σ_{n-1}) represented: i) the LE_A and 3 x LE_{TM} (concerns superior possible value of HLE-related LE_{TM} - 1 described above) values less than 0.05 with respect to any aa probability and any length range used in ELEMS(RDA), respectively, and ii) low SSD derived differences (analogues of arithmetical SSD) 8.6 x 10^{-3} and 7.5 x 10^{-3} corresponding to Q_c and Q_{cI}, respectively. The differences between logarithms of Q_c and Q_{cI} described above and logarithmic geometrical means (enumerated in height extents from seven to any value from 7 + MIN(U) = 13 to maximal height extent 26) were compared with confidence intervals for mean values (**CI**) (Zvárová, 2001). These differences exceeded half of the corresponding CI extents only in the ranges of heights higher than twenty two and twenty four, respectively, whereas any difference was not critical with respect to the compared whole CI extents.

WP3.4. Some formulas necessary for processing of data presented in Table WPT2 and subsections WP3.2.2 and WP3.2.3

WP3.4.1 Model BLAST sequence similarities. All BLAST searches usually yield low frequencies of minimal similarities keeping original query length (**MSQL**). To generate limiting similarities with the help of BLAST search, we used a two step procedure including selection of MSQL from search results and model substitution similar (BZ) or identical aa (BB) of selected MSQL by pair BX without similarity. Since MSQL have to be close to limits ten and 20 000 in case of Standard Protein BLAST (**SPrB**; BLOSUM62 matrix) and Search nearly exact matches (**SNEM**; PAM30 matrix), respectively, we can also construct Expect values of artificially modified MSQL:

$$E_m = E_o \times e^{\lambda \times \Delta S}, \qquad \text{(WPF27)}$$

where E_m, E_o are modified and original Expect values, λ is constant, ΔS is defined as $\Delta S = S_o - S_m$, where S_o is the original score of modified aa similarity, and S_m is the score of modified aa.

WP3.4.2 Calculation of q values related to different aa similarities out of mutation relationship. To compare ELEMS(RDA) and BLAST we have to approximate here BLAST positivities not yet included in our approach, i.e. the correlated aa pairs which are not convertible via single triplet mutation. An approximate corresponding relationship may follow from BLAST formula evaluating the Expect value:

$$q_{BZ} = (q_B \times q_Z)^{1/2} \times e^{\lambda \times (1/2 \times (S_{BB} + S_{ZZ}) - S_{BZ})}, \qquad \text{(WPF28)}$$

where $q_B \times q_Z$ are q values of compared aa, S_{BB}, S_{ZZ} are score identities of aa B and Z, and S_{BZ} is the score of unexplained positive matrix linkages between aa B and Z.

WP4 Frequently occurring terms and abbreviations

aa - amino acid residue(s); a_A - aa probability defined according to the used ELEMS version; a_{GM} - geometrical mean of tmaa probabilities; BC - backward comparison described in WP2.1; b_{TM} - probability of aa similarities in a given arrangement of sequence block with or without gaps; c_{TM} - probability of chain formed only by similar/identical aa (for details see WP2.1); ELEMS - evaluation of sequence similarities with the help of length equivalent measures; ELEMS(RDA) - the first version of ELEMS (for details see WP3.1.1-2); HLE - horizontal length equivalent(s) (see WP2.2); INF – inferior, meaning unachievable infimum; LE_A, LE_{TM} - VLE representing individual MTC, or mean model twin block height(s) related to aa similarities/identities, respectively; MSA - multiple sequence alignment(s); MTC - model twin column(s); SL - specific limit(s) defined in WP2.2; SIM(ALL) - overall block-chain similarity; TM - template motif(s); tmaa - template motif aa; VLE - vertical length equivalent(s); WPT1 - the table displayed in the associated web page; in bold - sites of the first local explanation of abbreviations.

WP5. References

Altschul, S.F., 1991. Amino acid substitution matrices from an information theoretic perspective. J. Mol. Biol. 219, 555-565.

Altschul, S. F., Bundschuh, R., Olsen, R., Hwa, T., 2001. The estimation of statistical parameters for local alignment score distributions. Nucleic Acids Res. 29, 351-361.

Baier, W., 1972. Theoretical approaches in analysis of systems, in: Drášil V. (Ed.), Biophysics, Academia, Praha, pp. 41-90.

Jura, P, 2003. Fuzzy systems, in: Vavřín P (Ed.), Fundamentals of fuzzy logics for regulation and modeling. Vitium. Brno, pp. 62-74.

Komenda, S., 1997. Probability as a fundament of statistics, in: Lepka I. (Ed.), Biometrics, University of Palacky, Olomouc, pp. 13-71.

Kubrycht, J., Borecký, J., 1998. Matrix formalization for simple approximate sequence comparison. Immunol. Zprav. 27/No. 3, 21-27.

Kubrycht, J., Borecký, J., Sigler, K., 2002. Sequence similarities of protein kinase peptide substrates and inhibitors: comparison of their primary structures with immunoglobulin repeats. Folia Microbiol. 47, 319-358.

Kubrycht, J., Borecký, J., Souèek, P., Ježek, P., 2004. Sequence similarities of protein kinase substrates and inhibitors with immunoglobulins and model immunoglobulin homologue: cell adhesion molecule from the living fossil sponge *Geodia cydonium*. Mapping of coherent database similarities and implications for evolution of CDR1 and hypermutation. Folia Microbiol. 49, 219-246.

Kubrycht, J., Sigler, K., Růžička, M., Souček, P., Borecký, J, Ježek, P., 2006. Ancient beginnings of immunoglobulin hypermutation. J. Mol. Evol. 63, 691-706.

Lepš, J., 1996. Biostatistics, University of Southern Bohemia, České Budějovice.

Olsen, R., Bundschuh, R., Hwa, T., 1999. Rapid assessment of extremal statistics for gapped local alignment. Proc. Int. Conf. Intell. Syst. Mol. Biol. 7, 211-222.

Perez-Bercoff, A., Koch, J., Burglin, T.R., 2006. LogoBar: bar graph visualization of protein logos with gaps. Bioinformatics 22, 112-114.

Poleksic, A., Danzer, J.F., Hambly, K., Debe, D.A., 2005. Convergent Island Statistics: a fast method for determining local alignment score significance. Bioinformatics 21, 2827-2831.

Schneider, T.D., Stephens, R.M., 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18, 6097-6100.

Tateno, Y., Takezaki, N., Nei, M.: Relative efficiencïes of the maximum-likelihood, neighbor-joining, and maximum parsinomy methods when substitution rate varies with site. Mol. Biol. Evol. 11, 261-267.

Zhang, J.I., 2005. Bayesian learning. http://merlin.cs.www.edu/~Zhang/cs571/notes /lec06.pdf

Zimmer, R., 2004. Vorlesung algorithmische bioinformatik II. XI: Orthodox and Bayesian modeling, HMMs. http://cgi.bio.ifi.lmu.de/lehre/WS2004/VLG_Algo_2.

Zvárová, J., 2001. Biomedical statistics. I Elements of statistics for biomedical branches. Karolinum, Prague.

.