Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis

July 10. 2006

This document is divided into two. The first chapter contains errors and errata, and the second chapter proposals for making some text more clear.

We thanks several people for sending us information about errors and errata, and confusions: Tom E. Gundersen, Ragnhild H. Be, Saywan Pasha, Knut A. Syed, and especially Lynda Ellis and Sara Vetter from University of Minnesota, Josef Thingnes from University of Oslo, and Eivind Coward from University of Bergen.

Note that there was a reprint Autumn 2004, where most of the reported errors for the first printing were corrected (but unfortunately some new introduced). In the listing we specify if the errors refer to the original *O* or reprint *R*, or both.

Errors and errata

Appendix A - Basics in Mathematics, Probability and Algorithms

- pg 315 *OR* - faculty should be factorial.
- pg 317 *O* - italicize "true" in the third assertion.
- pg 318 *OR* - In A.4.2, value space is normally called range.
- pg 321 *O* - Table A.3 title, change "faculty" to "factorial"
- pg 321 *O* - second complete paragraph, line 2, change '(for $k > 1$' to '(for $k > 1$).'

Appendix B - Introduction to Molecular Biology

- pg 324 *O* - first full paragraph, line 9 change "pyremidine" to "pyrimidine"
- pg 326 *O* - first full paragraph, line 2 change "pyremidine" to "pyrimidine"

Chapter 1 - Pairwise Global Alignment of Sequences

- pg 7 *OR* - There are wrong spacing in the alignments.
- pg 10 *OR* - Algorithm 1.1 The first sentences should be
  for $i := 0$ **to** m **do** $H_{i,0} := -ig$ **end**
  for $j := 0$ **to** n **do** $H_{0,j} := -jg$ **end**

- pg 11 *OR* - Figure 1.1 should be Table 1.1.
- pg 12 *OR* - Algorithm 1.2 Many of the assignment are expressed with "=" instead of with ": ="
- pg 13 *O* - minus signs are omitted in Figure 1.3a row 2 (starts with V). Figure 1.3b is correct.

Chapter 2 - Pairwise Local Alignment and Database Search

- pg 27 *OR* - Line four in Section 2.2, $q(i) = d(j)$ should be $q_i = d_j$
- pg 28 *OR* - There is a conflict between the text and Figure 2.1. The text and the algorithm places the dot in the first position of the window, but in the figure it is placed in the middle position.
- pg 30 *R* - Algorithm 2.1 change "if sum(..." to "if sim(..." (for similarity)
- pg 30 *OR* - Algorithm 2.1 at "if sim(..", add a missing ")"
- pg 34 *O* line 2, change "now" to "no".
- pg 42 *OR* - Some assignments with "=" instead of ":="
- pg 43 *OR* - line 12, an ")" is missing

Chapter 3 - Statistical Analysis

- pg 49 *OR* - Figure 3.1.
  The x-axis of the figure should have minus signs in front of the 2 and 1 on the left side of 0.
- pg 52 *OR* - Normal distribution should be normal distribution.
- pg 53 *OR* - first line middle value should be mean value.
- pg 55 *OR* - Title 3.3.1 should be "The S_M score has an extreme value distribution".
- pg 61 *OR* - an extra ")" in the first line after the formula
- pg 63 *O* - exercise 1(b)
  change "u (the ratio of the modal to the characteristic value)" to "u (the modal, or characteristic, value)"
  change "lambda (the variance measure and decay constant)" to "lambda (the variance measure or decay constant)"

Chapter 4 - Multiple Global alignment and Phylogenetic Trees

- pg 77 *OR* - $Tr0ot(10)$ should be $Troot(10)$
- pg 86 *OR* - In Figure 4.16, (0) for the tree T should be (1).
- pg 88 *OR* - In Equation 4.12, is an extra Z
- pg 98 *OR* - Exercise 7. By definition an ultrametrix tree is additive, but show that Equation 4.10 follows from Equation 4.11.

Chapter 5 - Scoring Matrices

- pg 107 *O* - First bullet point, change: indicate high mutability; to indicates high mutability;
- pg 107 *O* - Section 5.2.3, second line change: the only changes to: the only change.
- pg 118 *OR* - In legend to Table 5.4, the expected value for BLOSUM 62 should be -0.52.
- pg 119 *OR* - In Equation (5.8) the argument of ln should be the whole fraction $\frac{h_{ab}}{p_a p_b}$.

Chapter 6 - Profiles

- pg 125 *OR* - Table 6.1. This profile is not the correct profile for the alignment in Figure 4.1, with PAM250. You can yourself construct the profile from http://eta.embl-heidelberg.de:8000/profw/
  for the alignment in Figure 4.1

```
Clustal X (1.64b) multiple sequence alignment
XENLA1  YPKVKRDMEQALVSGPQD------NELDG--MQLQPQ--EYQKMKRGIVEQ
XENLA2  YPKIKRDIEQAQVNGPQD------NELDG--MQFQPQ--EYQKMKRGIVEQ
MOUSE1  TPKSRREVEDPQVEQLEL------GGSP---GDLQTLALEVARQKRGIVDQ
RAT1    TPKSRREVEDPQVPQLEL------GGGPEA-GDLQTLALEVARQKRGIVDQ
MOUSE2  TPMSRREVEDPQVAQLEL------GGGPGA-GDLQTLALEVAQQKRGIVDQ
RAT2    TPMSRREVEDPQVAQLEL------GGGPGA-GDLQTLALEVARQKRGIVDQ
CRILO   TPKSRRGVEDPQVAQLEL------GGGPGA-DDLQTLALEVAQQKRGIVDQ
RABIT   TPKSRREVEELQVGQAEL------GGGPGA-GGLQPSALELALQKRGIVEQ
BOVIN   TPKARREVEGPQVGALEL------AGGPG------AGGLEGPPQKRGIVEQ
SHEEP   TPKARREVEGPQVGALEL------AGGPG------AGGLEGPPQKRGIVEQ
PIG     TPKARREAENPQAGAVEL------GGGLGG---LQALALEGPPQKRGIVEQ
CANFA   TPKARREVEDLQVRDVEL------AGAPGE-GGLQPLALEGALQKRGIVEQ
HUMAN   TPKTRREAEDLQVGQVEL------GGGPGA-GSLQPLALEGSLQKRGIVEQ
PANTR   TPKTRREAEDLQVGQVEL------GGGPGA-GSLQPLALEGSLQKRGIVEQ
CERAE   TPKTRREAEDPQVGQVEL------GGGPGA-GSLQPLALEGSLQKRGIVEQ
AOTTR   APKTRREAEDLQVGQVEL------GGGSIT-GSLPP--LEGPMQKRGVVDQ
CAVPO   IPKDRRELEDPQVEQTEL------GMGLGA-GGLQPLALEMALQKRGIVDQ
CHICK   SPKARRDVEQP-LVSSPL------RGEAGV-LPFQQE--EYEKVKRGIVEQ
ORENI   NPR--RDVDPLLGFLPPKAGGAVVQGGEN---EVTFKDQMEMMVKRGIVEE
VERMO   TPK--RDVDPLLGFLPAKSGGAAAGG-ENEVAEFAFKDQMEMMVKRGIVEQ
BRARE   NPK--RDVEPLLGFLPPK------SAQETEVADFAFKDHAELIRKRGIVEQ
ONCKE   TPK--RDVDPLIGFLSPK------SAKENE--EYPFKDQTEMMVKRGIVEQ
         *   *  :                                     ***:*::
```

- pg 129 *OR* - line six. “for small m_r” should be “for small T_{ra}”

Chapter 7 - Sequence Patterns

- pg 146 *OR* - The legend of Figure 7.1 should be “The entropy for a binary alphabet.”

Chapter 8 - Structures and Structure Description

- pg 174 *OR* - In Figure 8.7 C_{i1} should be C_{i-1}

- pg 177 *O* - Section 8.5.4, In the equation for antiparallell bonding, $j+2$ should be $j-2$.

Chapter 9 - Superposition and Dynamic Programming

- pg 194 *OR* - For a matrix to be a rotation matrix it must, in addition to be orthogonal, it's determinant must also be equal to one.

- pg 194 *OR* - The matrix R in the Example is not a rotation matrix, it is not orthogonal, and its determinant is not $+1$. An example of a rotation matrix is $\begin{array}{c} \begin{array}{ccc} C_1 & C_2 & C_3 \end{array} \\ \left\{ \begin{array}{ccc} -\frac{1}{2} & -\frac{\sqrt{3}}{4} & \frac{3}{4} \\ \frac{\sqrt{3}}{2} & -\frac{1}{4} & \frac{\sqrt{3}}{4} \\ 0 & \frac{\sqrt{3}}{2} & \frac{1}{2} \end{array} \right\} \end{array}$.

- pg 195 *OR* - The point $\{1,-1,1\}$ rotated with this new rotation matrix becomes $\{\frac{1+\sqrt{3}}{4}, \frac{1+3\sqrt{3}}{4}, \frac{1-\sqrt{3}}{2}\}$

- pg 208 *O* - Exercise 2(b), the given matrix is not a rotation matrix (the determinant not equal to $+1$) To get a rotation matrix, change the sign of the first element in column one to pluss, and the sign of the second element in column one to minus.

- pg 208 *R* - Exercise 2. The rotation matrix is corrected in the reprint, but not the coordinates of the structures (to get reasonable results). Change the coordinates of A to be (0.5858, 1)(1.2929, -1.1213) (4.1213, 0.2929)(2, 3.8285). The new coordinates of a_1 should then be (-1,-1).

Chapter 11 - Clustering - Combining Local Similarities

- pg 237 *O* - Algorithm 11.3
  The algorithm is wrong, such that it produces a linear clustering. A heuristic algorithm for hierarchical clustering is

  Algorithm 11.3 Hierarchical clustering

  var
  $\mathcal{C}$ the set of current clusters
  $\mathcal{H}$ the pairs of consistent clusters in $\mathcal{C}$ that increases score

	by grouping them
$S(C)$	the score of a cluster C
C_1, C_2	current highest scoring pair of clusters in $\mathcal{H}$
$S(C_1, C_2)$	the score of grouping (C_1, C_2)

proc

$incr(C_i, C_j) = cons(C_i, C_j)$ **and** $[S(C_i, C_j) > \max(S(C_i), S(C_j))]$
$incr(C_i, C_j)$ is true if C_i and C_j are consistent,
and the score increases by grouping them

begin
$\mathcal{C}$:=*set of clusters, each seed match being a cluster*
$\mathcal{H} := \{(C_i, C_j) | C_i \in \mathcal{C}$ **and** $C_j \in \mathcal{C}$ **and** $incr(C_i, C_j)\}$
determine C_1, C_2
while $\mathcal{H} \neq \emptyset$ **do**
$C_B :=group(C_1, C_2)$ the highest scoring pair in $\mathcal{H}$
$\mathcal{C} := (\mathcal{C} - \{C_1, C_2\}) \cup \{C_B\}$
update $\mathcal{H}$ *and determine* C_1, C_2
end
end $\mathcal{C}$ now contains the found clusters,
and no pairs will increase the score

- pg 241 *OR* - In the Example, L_k should also contain (14,11)
- pg 242 *OR* - Algorithm 11.4 has the same eror as in Algorithm 11.3 in the original printing.

Chapter 15 - Structure prediction: Threading

- pg 302 *O* - Just after Equation (15.1) buried states should be burial states.

Clearifying text

Chapter 1 - Pairwise Global Alignment of Sequences

- Small triangles indicates the end of Chapter examples.
- pg 7-8. Symbol here means residue.
- pg 14 $\forall r$ means for all r in the given range
- pg 16 The first sentence is not true for linear gap penalty.
- pg 20. In this context, *string* is the same as sequence.
- pg 23. Exercise 1.2.d. Note that in this exercise shall end gaps not be penalised.

Chapter 2 - Pairwise Local Alignment and Database Search

- pg 44 Question 5, Equation 1.2 is in Section 1.7.

Chapter 3 - Statistical Analysis

- pg 64 Exercise 5. Remember that exp(x) = e to the x and ln(e to the x) = x.

Chapter 4 - Multiple Global Alignment and Phylogenetic Trees

- pg 80. Note that the informative columns (3, 5, 7) are not the same as those in Figure 4.10 (columns 3, 4, 5).

- pg 81-84. Note how the *distance* here is used between internal nodes. It is not the distance between the nodes, but between the *subtrees* where the nodes are the roots.

Chapter 5 - Scoring Matrices

- pg 121. Question 7b(i) change: how many residues that have not changed; to: the number of residues that have not changed;

Chapter 6 - Profiles

- pg 137. Note that in the equation in Section 6.4.3, the maximum is taken over all k in the interval $0 \leq k \leq j-1$, i and j are fixed.

Chapter 7 - Sequence Patterns

- pg 145. In the example, the number of patterns are found as

  $$9 \cdot (20^2 + 20^3) \cdot (20^2 + 20^3) = 9 \cdot 20^2(1+20) \cdot 20^2(1+20) = 9 \cdot 20^4 \cdot 21^2.$$

- pg 161. Question 3, part (b): change: “and align it to the two other sequences” to:
  “and align it to each of the two other sequences. Score each alignment”.

- pg 161. Question 3, part (c): change: “By using the found scores, decide which patterns/sequences to align next, and perform the aligning”. to: “Based on the scores, decide which patterns/sequences to align next, and perform the alignment”.

Chapter 8 - Structures and Structure Descriptions

- pg 171. The Lagrange theorem is used to find the distance from each point to the centroid.