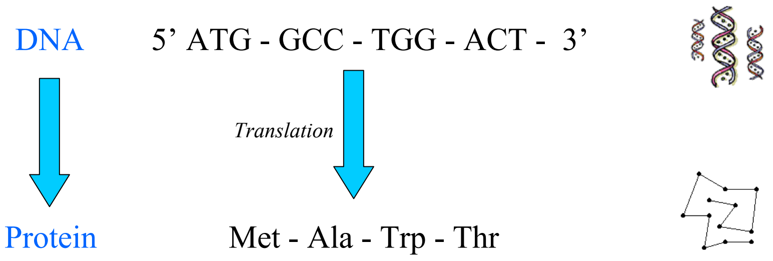
An overview of the computational analysis of biological sequences

Prof. S Subbiah
Department of Applied Physics
Stanford University
&
Bioinformatics Centre
National University of Singapore

Introduction

- Basics

Biological sequences and their meaning



DNA

5’ ATG - GCC - TGG - ACT - 3’

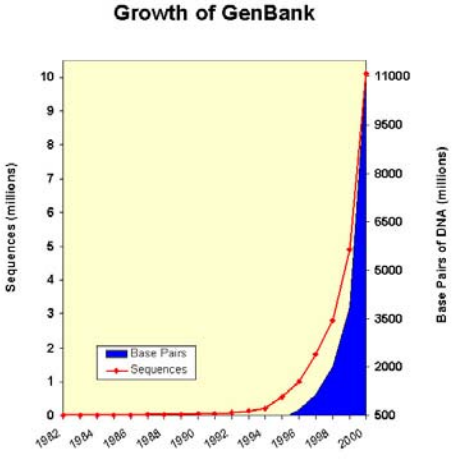
Translation

Protein

Met - Ala - Trp - Thr

Exponential growth in known sequences

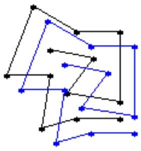
- doubles approximately every 14 months.
- contains more than 11 billion bases from over 100,000 species



Growth of GenBank

Sequence & structure alignments

```
            10        20        30        40        50
    MARYRCCLTHSGSRCRRRRRRRCRRRRRRFGRRRRRRVCCRRYTVIRCTRQ
    ::::::: .:: :::: v^:: :::::::   :::: ::::::::::X :
SPSIMARYRCCRSHSRSRCRPRRRR-CRRRRRRCCPRRRRAVCCRRYTVIRCRRC
       10        20         30        40        50
```



Topics to be covered

- Basis for sequence alignment (1 hr of lectures)
  - Evolutionary
  - Structural
- Pairwise sequence alignment (2 hrs of lectures)
  - Why ?
  - How ?
  - Example
- Multiple sequence alignment (2 hrs of lectures)
  - Why ?
  - How ?
  - Example

Topics to be covered

- Searching sequence databases (2 hrs of lectures)
  - Why ?
  - How ?
  - Example.
- Other topics (not covered) (5 hrs of lectures)
  - Pattern /Motif searching
  - Gene structure prediction

Basis for sequence alignment

- Evolutionary
- Structural

Evolutionary basis of alignment

- Enable the researcher to determine if two sequences display sufficient similarity to justify the inference of homology.
- Similarity is an observable quantity that may be expressed as say %identity or some other measure.
- Homology is a conclusion drawn from this data that the two genes share a common evolutionary history.

Evolutionary basis of alignment

- Genes are either homologous or not homologous.
- There are no degrees of homology as are there in similarity.
- While it is presumed that the homologous sequences have diverged from a common ancestral sequence through iterative molecular changes we do not actually know what the ancestral sequence was.

Evolutionary basis of alignment

- Thus an alignment just reflects the **probable** evolutionary history of the two genes for the proteins.
- Residues that have aligned and are not identical represent substitutions.
- Regions in which the residues of one sequence correspond to nothing in the other would be interpreted as either an insertion/deletion. These regions are represented in an alignment as Gaps.

Evolutionary basis of alignment

- Certain regions are more conserved than others - crucial residues (structure/function)
- There may be certain regions conserved but not functionally related - historical reasons.
- Especially, from closely related species- have not had sufficient time to diverge.

Structural basis for alignment

- It is well-known that when two protein sequences have more than 20-30% identical residues aligned the corresponding 3-D structures are almost always structurally very similar.
- Overall folds are identical & structures differ in detail.
- Form often follows function. So sequence similarity by way of structural similarity implies similar function.
- So the sequence alignment is often an approximate predictor of the underlying 3-D structural alignment.

Caveat

- Computational predictions only suggest - to make a conclusive case further experimental tests must validate.
- Evolutionary relatedness must be confirmed either by - experimental evidence for evolutionary history or experimental establishment of similar function.
- For structural relatedness the 3-D structures must be experimentally determined and compared.

Pairwise sequence alignment

- Why ?
- How ?
- Example

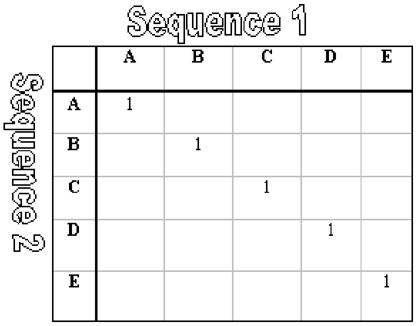
Why ?

Basis for other analyses

- Inference of protein's function
  - Conserved positions - functionally critical residues ?
- Structure prediction
  - Pattern of hydrophobicity - suggests secondary structure
  - Gap regions - suggest structural loops
  - Fold prediction, homology modelling
- Phylogeny
- Experimental Design
  - PCR primer design
  - Mutagenesis studies
- Detection of previously known motifs

How ?

Classic Needleman-Wunsch algorithm



Sequence 1

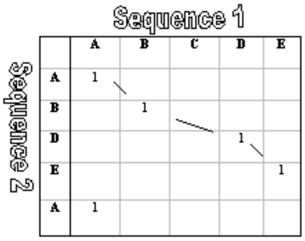
Sequence 2

	A	B	C	D	E
A	1				
B		1			
C			1		
D				1	
E					1

Sequence 1 A B C D E
Sequence 2 A B C D E

How ?

Classic Needleman-Wunsch algorithm



Sequence 1

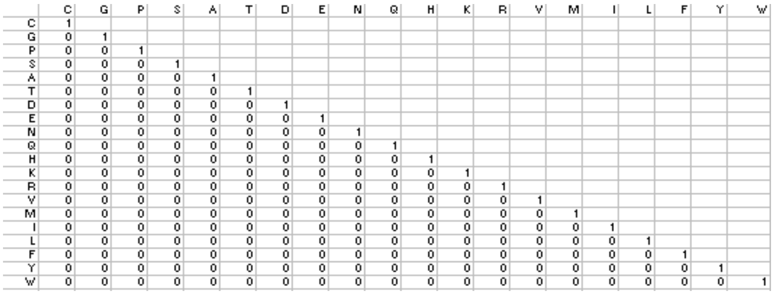
Sequence 2

	A	B	C	D	E
A	1				
B		1			
D				1	
E					1
A	1				

Sequence 1 A B C D E _
Sequence 2 A B _ D E A

Common scoring matrices

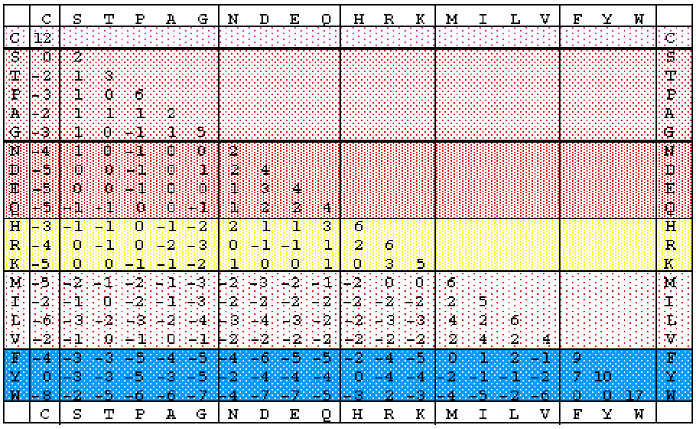
Based on identities



	C	G	P	S	A	T	D	E	N	Q	H	K	R	V	M	I	L	F	Y	W
C	1																			
G	0	1																		
P	0	0	1																	
S	0	0	0	1																
A	0	0	0	0	1															
T	0	0	0	0	0	1														
D	0	0	0	0	0	0	1													
E	0	0	0	0	0	0	0	1												
N	0	0	0	0	0	0	0	0	1											
Q	0	0	0	0	0	0	0	0	0	1										
H	0	0	0	0	0	0	0	0	0	0	1									
K	0	0	0	0	0	0	0	0	0	0	0	1								
R	0	0	0	0	0	0	0	0	0	0	0	0	1							
V	0	0	0	0	0	0	0	0	0	0	0	0	0	1						
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1					
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1				
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Common scoring matrices

Based on observed mutational rates (Dayhoff, 1970's)



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

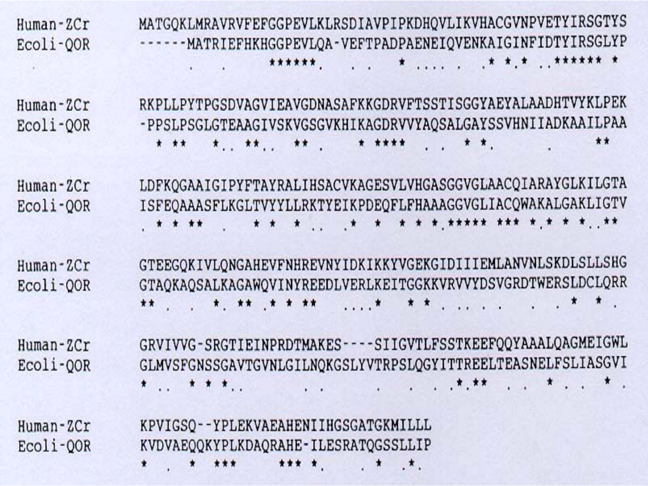
How ?

Other methods

Align	(Dayhoff, early 80’s)
Gap	(GCG package)
Smith-Waterman	(Smith, Waterman)
Bestfit	(GCG Package)
FASTA	(Pearson, Lipman)
BLAST	(Altschul, Lipman)
HMM methods	(Haussler, Eddy)
Etc.	

Example – Cautionary tale of Function Inference

The case of human eye lens protein and E.coli metabolic enzyme



```
Human-ZCr      MATGQKLMRAVRVFEFGGPEVLKLRSDIAVPIPKDHQVLIKVHACGVNPVETYIRSGTYS
Ecoli-QOR      ------MATRIEFHKHGGPEVLQA-VEFTPADPAENEIQVENKAIGINFIDTYIRSGLYP
  .                  .   .     ******.   . .   * .... .  .* *.* ..****** *

Human-ZCr      RKPLLPYTPGSDVAGVIEAVGDNASAFKKGDRVFTSSTISGGYAEYALAADHTVYKLPEK
Ecoli-QOR      -PPSLPSGLGTEAAGIVSKVGSGVKHIKAGDRVVYAQSALGAYSSVHNIIADKAAILPAA
                 * **   *.. **..  ** .    * ****  . .  * *.      .     **

Human-ZCr      LDFKQGAAIGIPYFTAYRALIHSACVKAGESVLVHGASGGVGLAACQIARAYGLKILGTA
Ecoli-QOR      ISFEQAAASFLKGLTVYYLLRKTYEIKPDEQFLFHAAAGGVGLIACQWAKALGAKLIGTV
               . * * **  .   * *  * ..  .*  *  * * *.***** *** *.* * *..**

Human-ZCr      GTEEGQKIVLQNGAHEVFNHREVNYIDKIKKYVGEKGIDIIIEMLANVNLSKDLSLLSHG
Ecoli-QOR      GTAQKAQSALKAGAWQVINYREEDLVERLKEITGGKKVRVVYDSVGRDTWERSLDCLQRR
               ** .  .  *. ** .* * **   ....*   * * . .. . .   .  . *  * .

Human-ZCr      GRVIVVG-SRGTIEINPRDTMAKES----SIIGVTLFSSTKEEFQQYAAALQAGMEIGWL
Ecoli-QOR      GLMVSFGNSSGAVTGVNLGILNQKGSLYVTRPSLQGYITTREELTEASNELFSLIASGVI
               * ..  * * *..       . .      .   .  . .*.**  . .  * . .  * .

Human-ZCr      KPVIGSQ--YPLEKVAEAHENIIHGSGATGKMILLL
Ecoli-QOR      KVDVAEQQKYPLKDAQRAHE-ILESRATQGSSLLIP
               *  .  *  ***     *** *.    . *  .*.
```

Example – Cautionary tale of Function Inference

- Appear to share a high degree of similarity.
- Should have similar biological function.
- Hypothetical statement.
- Crystalline: lens matrix of vertebrate eye.
  E.coli metabolic enzyme - quinone oxido reductase
- Function has changed during the course of evolution.
- BE CAREFUL !!

Multiple sequence alignment

- Why ?
- How ?
- Example

Why ?

Basis for other analyses:

- All the ones previously discussed for pairwise but with more subtle information.
- Creation of sequence profiles for searching (PROSITE, PRINTS, PFAM, BLOCKS)
- Deduction of sequence motifs.
- Useful for analysing protein family relationships.
- A convenient backdrop for annotation summarising information results of various sequence analysis.

How ?

Progressive alignment approach – e.g.TULLA (Subbiah, 1984)

- Create a tree by comparing the most similar sequences step by step.
- The two most similar sequences are aligned and then the next two sequences are aligned that are most similar.
- Re-adjust the gaps so that the alignment is maximum and the gaps are least.
- Heuristic approach, in theory not always optimal.

Add new sequences to an existing
Alignment

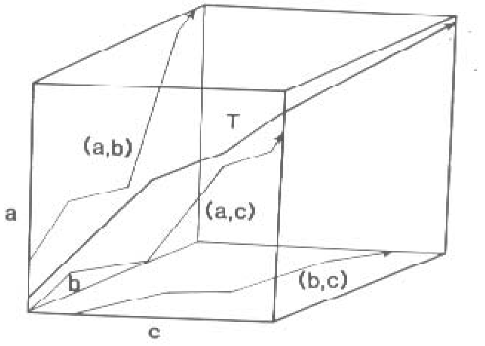
1. One new sequence

2. A set of new sequences, added one at a time

Use sequence weights to ensure that new sequence is aligned to most closely related sequence.

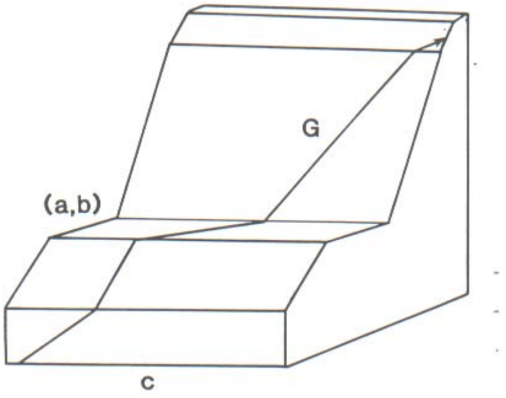
How ?

Brute Force multi-dimensional Needleman-Wunsch is computationally unrealistic



How ?

TULLA - Progressive Needleman-Wunsch of locked-sets



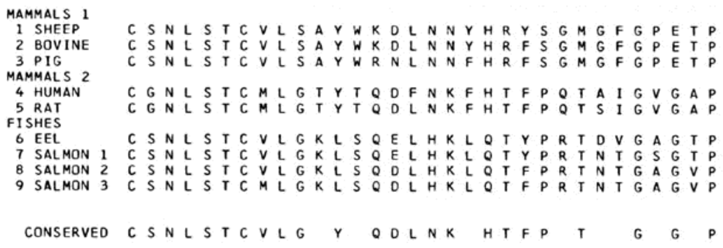
How ?

Other methods

- Feng & Doolittle
- Clustal
- ClustalW
- PileUp
- SAGA, Genetic Algorithms
- Etc.

Example – Creation of a family alignment

The case of the Calcitonin proteins



```
MAMMALS 1
 1 SHEEP      C S N L S T C V L S A Y W K D L N N Y H R Y S G M G F G P E T P
 2 BOVINE     C S N L S T C V L S A Y W K D L N N Y H R F S G M G F G P E T P
 3 PIG        C S N L S T C V L S A Y W R N L N N F H R F S G M G F G P E T P
MAMMALS 2
 4 HUMAN      C G N L S T C M L G T Y T Q D F N K F H T F P Q T A I G V G A P
 5 RAT        C G N L S T C M L G T Y T Q D L N K F H T F P Q T S I G V G A P
FISHES
 6 EEL        C S N L S T C V L G K L S Q E L H K L Q T Y P R T D V G A G T P
 7 SALMON 1   C S N L S T C V L G K L S Q E L H K L Q T Y P R T N T G S G T P
 8 SALMON 2   C S N L S T C V L G K L S Q D L H K L Q T F P R T N T G A G V P
 9 SALMON 3   C S N L S T C M L G K L S Q D L H K L Q T F P R T N T G A G V P

  CONSERVED   C S N L S T C V L G   Y   Q D L N K   H T F P   T     G   G   P
```

Example - Generation of Profiles

The case of DNA recognition elements

Multiple
Sequence alignment

```
A   A   T   T   -   G   G   A   A   C
A   A   T   T   -   -   G   A   A   C
A   A   T   T   T   G   G   A   A   C
-   A   T   T   T   G   G   A   -   C
A   A   T   T   -   G   G   A   T   C
-   A   T   -   -   -   G   A   A   C
-   A   T   -   -   -   -   A   A   C
A   A   T   T   -   -   G   A   A   -
-   A   T   -   -   -   -   A   T   C
A   A   T   -   -   -   -   A   T   C
```

Profile derived
from the alignment

```
A  0.6  1.0  0.0  0.0  0.0  0.0  0.0  1.0  0.6  0.0

C  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.9

G  0.0  0.0  0.0  0.0  0.0  0.4  0.7  0.0  0.0  0.0

T  0.0  0.0  1.0  0.6  0.2  0.0  0.0  0.0  0.3  0.0
```

Searching large sequence databases

- Why ?
- How ?
- Example

Why ?

- Is there any protein sequence that is similar to mine ?
- Is this gene known in any other species ?
- Has someone already identified this sequence ?
- Can we guess function before tedious experimentation ?

How ?

Progressive alignment approach – e.g.TULLA (Subbiah, 1984)

- Assume n = 3 sequences to be aligned – *a* , *b* & *c*.
- Align all pairs of sequences – *(a,b), (b,c)* & *(a,c)* – by pairwise NW and pick the best aligned/most related pair, say *(a,b)*.
- Note, the *(a,b)* alignment may have gaps/insertions in a relative to *b*. At all subsequent steps we keep this (*a,b*) alignment internally "locked" – no further gaps/insertions introduced in *a* relative to *b*.
- Now align this best locked pair *(a,b)* against the next sequence *c* and obtain a 3-way alignment, *G*. Note that gaps are introduced either in sequence *c* or to both sequences *a* & *b* simultaneously.
- When n > 3, repeat as necessary using increasingly larger locket-sets.
- Heuristic method, no guarantee of mathematical optmimum.
- *G* is not equal to *T* , but often good enough approximation.

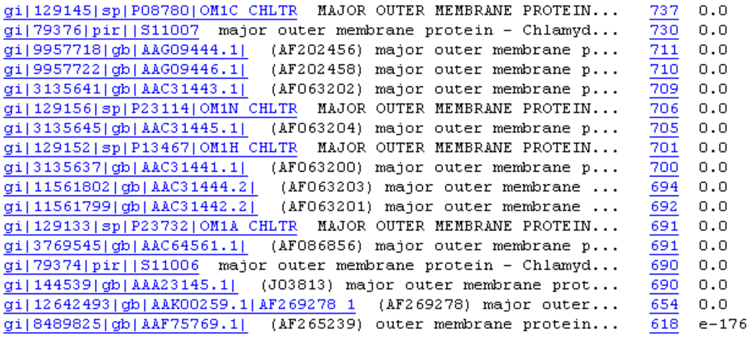
How ?

Other methods

- BLAST
- FASTA
- PSI-BLAST
- PHI BLAST
- Etc.

Example – Finding related proteins in databases

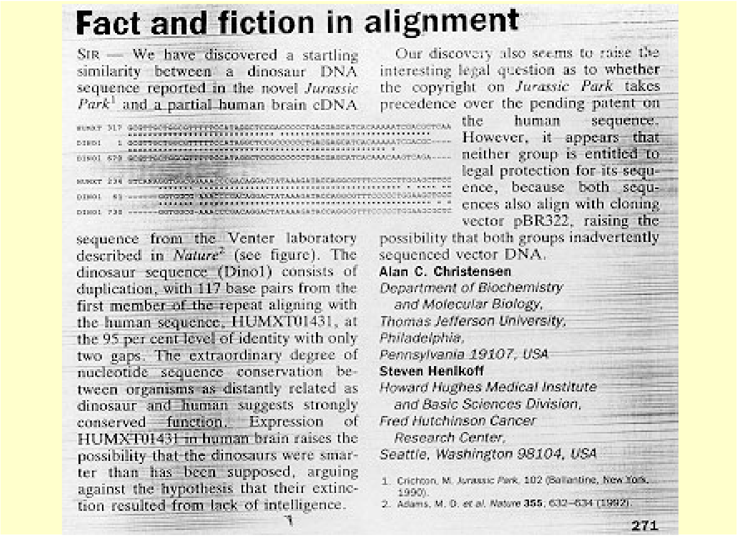
The case of major outer membrane protein from Chlamydia



```
gi|129145|sp|P08780|OM1C CHLTR  MAJOR OUTER MEMBRANE PROTEIN...    737  0.0
gi|79376|pir||S11007  major outer membrane protein - Chlamyd...    730  0.0
gi|9957718|gb|AAG09444.1|  (AF202456) major outer membrane p...    711  0.0
gi|9957722|gb|AAG09446.1|  (AF202458) major outer membrane p...    710  0.0
gi|3135641|gb|AAC31443.1|  (AF063202) major outer membrane p...    709  0.0
gi|129156|sp|P23114|OM1N CHLTR  MAJOR OUTER MEMBRANE PROTEIN...    706  0.0
gi|3135645|gb|AAC31445.1|  (AF063204) major outer membrane p...    705  0.0
gi|129152|sp|P13467|OM1H CHLTR  MAJOR OUTER MEMBRANE PROTEIN...    701  0.0
gi|3135637|gb|AAC31441.1|  (AF063200) major outer membrane p...    700  0.0
gi|11561802|gb|AAC31444.2|  (AF063203) major outer membrane ...    694  0.0
gi|11561799|gb|AAC31442.2|  (AF063201) major outer membrane ...    692  0.0
gi|129133|sp|P23732|OM1A CHLTR  MAJOR OUTER MEMBRANE PROTEIN...    691  0.0
gi|3769545|gb|AAC64561.1|  (AF086856) major outer membrane p...    691  0.0
gi|79374|pir||S11006  major outer membrane protein - Chlamyd...    690  0.0
gi|144539|gb|AAA23145.1|  (J03813) major outer membrane prot...    690  0.0
gi|12642493|gb|AAK00259.1|AF269278 1  (AF269278) major outer...    654  0.0
gi|8489825|gb|AAF75769.1|  (AF265239) outer membrane protein...    618  e-176
```

Concluding Example – An amusing tale

The intelligence of dinosaurs, old and new



Fact and fiction in alignment

SIR — We have discovered a startling similarity between a dinosaur DNA sequence reported in the novel *Jurassic Park*[1] and a partial human brain cDNA sequence from the Venter laboratory described in *Nature*[2] (see figure). The dinosaur sequence (Dino1) consists of duplication, with 117 base pairs from the first member of the repeat aligning with the human sequence, HUMXT01431, at the 95 per cent level of identity with only two gaps. The extraordinary degree of nucleotide sequence conservation between organisms as distantly related as dinosaur and human suggests strongly conserved function. Expression of HUMXT01431 in human brain raises the possibility that the dinosaurs were smarter than has been supposed, arguing against the hypothesis that their extinction resulted from lack of intelligence.

Our discovery also seems to raise the interesting legal question as to whether the copyright on *Jurassic Park* takes precedence over the pending patent on the human sequence. However, it appears that neither group is entitled to legal protection for its sequence, because both sequences also align with cloning vector pBR322, raising the possibility that both groups inadvertently sequenced vector DNA.

Alan C. Christensen
Department of Biochemistry
and Molecular Biology,
Thomas Jefferson University,
Philadelphia,
Pennsylvania 19107, USA
Steven Henikoff
Howard Hughes Medical Institute
and Basic Sciences Division,
Fred Hutchinson Cancer
Research Center,
Seattle, Washington 98104, USA

[1] Crichton, M. *Jurassic Park*, 102 (Ballantine, New York, 1990).
[2] Adams, M. D. *et al. Nature* **355**, 632–634 (1992).