Tin Sinh học
Bioinformatics

Chương 4
1. BLAST

TS. Nguyễn Hồng Quang
Khoa Kỹ thuật máy tính
Leader of Bioinformatics Group, BK.AI center
Trường Công nghệ thông tin và Truyền thông
Trường Đại học Bách Khoa Hà Nội

Tài liệu tham khảo

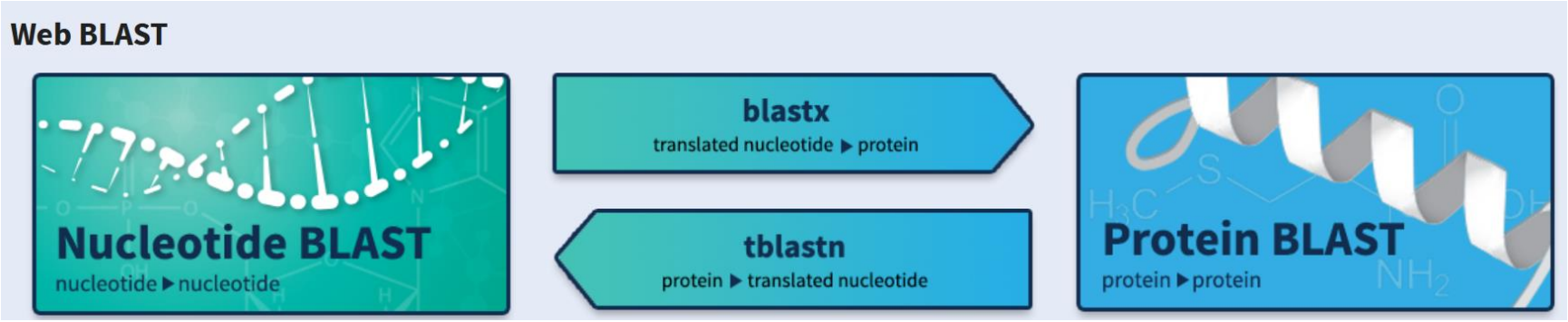
- Nicholas James Provart, Bioinformatic Methods I, Coursera, University of Toronto, 2021.

Nội dung

- Ma trận thay thế (Substitution matrices)
- Giải thuật BLAST
- Độ đo đánh giá kết quả tìm được

BLAST

- Để xác định các đặc điểm, chức năng của một trình tự (DNA, RNA, protein) mới:
  - Xác định các trình tự tương tự trong cơ sở dữ liệu
  - => các trình tự tương đồng
  - => giống nhau về chức năng của gen



Liên kết trình tự

- Sequence alignments
- Dấu . : các amino acids có các tính chất hóa lý tương tự nhau
- Để thực hiện "Sequence alignments" cần "Substitution matrices" (ma trận thay thế)

```
Sequence 1:  HEAGAWGHEE

Sequence 2:  PAWHEAE


Sequence 1: HEAGAWGHE-E
                  .  ++  ++  +
Sequence 2: --P-AW-HEAE
```

Substitution matrices

- Mô hình hóa được sự thay đổi trình tự tiến hóa theo thời gian
- Cần đánh giá:
  - substitution biases
  - mutational saturation

	A	R	N	D	C
A	2				
R	-2	6			
N	0	0	2		
D	0	-1	2	4	
C	-2	-4	-4	-5	12

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

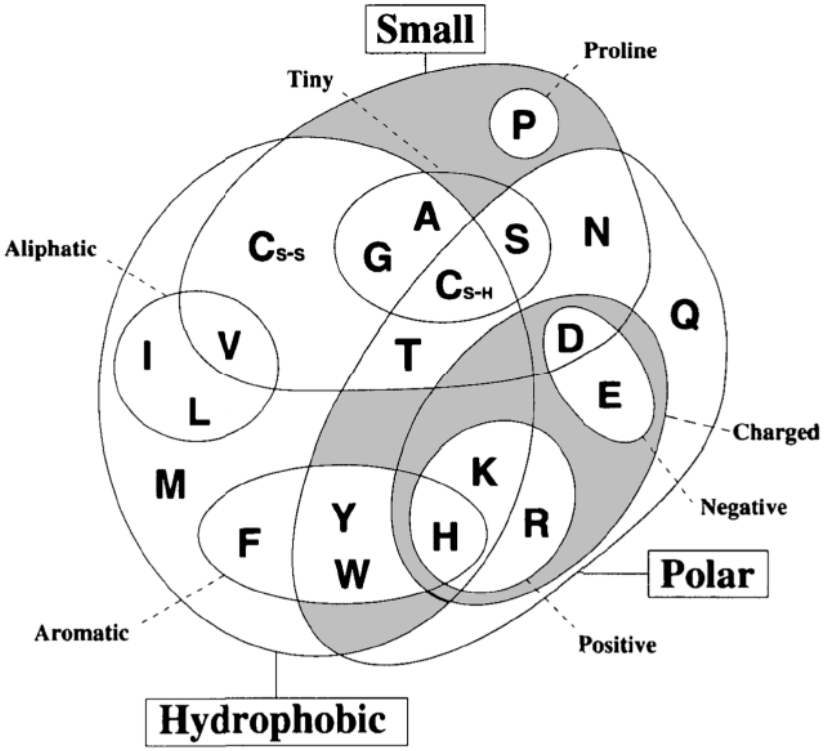
Substitution matrices – substitution biases

Amino acids biochemical properties	nonpolar	polar	basic	acidic

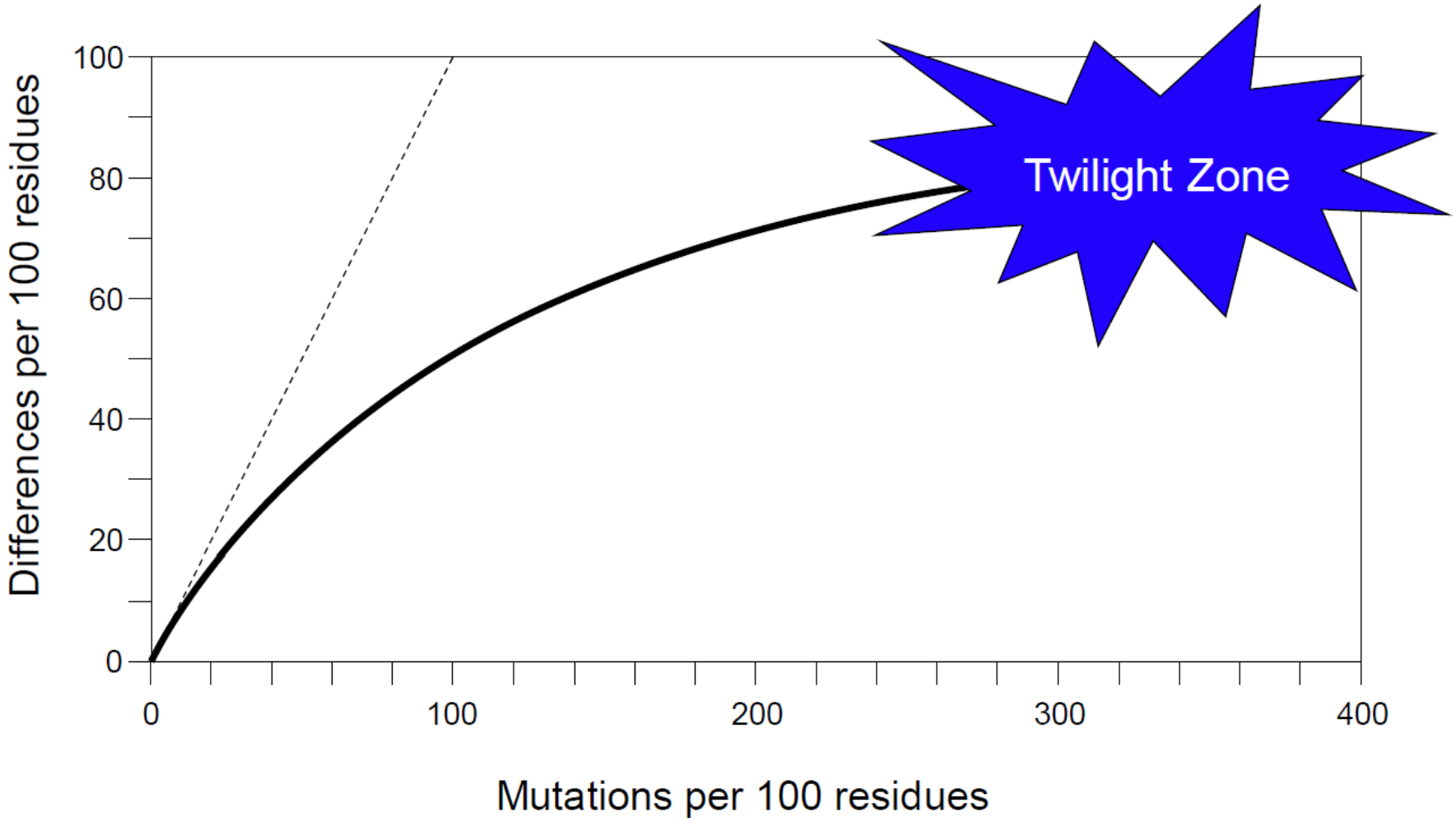
Termination: stop codon

Standard genetic code

1st base	2nd base: T		C		A		G		3rd base
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*)[B]	TGA	Stop (*Opal*)[B]	A
	TTG[A]		TCG		TAG	Stop (*Amber*)[B]	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG[A]		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G



Substitution matrices – mutational saturation



Các yêu cầu của ma trận thay thế

- Mô hình hóa được sự thay đổi trình tự tiến hóa theo thời gian
- Hỗ trợ matching các amino acids giống nhau và có liên quan với nhau
- Penalize các amino acids hoặc gaps "poorly matched"
- Có tính đến sự dư thừa của một số loại amino acids trong protein, ví dụ như alanine

Ma trận thay thế amino acids

- PAM = Point Accepted Mutations: sử dụng các alignment giữa các trình tự có liên quan gần nhau

PAM250

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

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Ma trận thay thế amino acids

- BLOSUM = Blocks Amino Acid Substitution Matrices:
- Sử dụng ungapped multiple alignments của các vùng bảo tồn ngắn (3-60aa) trong CSDL BLOCKS của các protein có mối quan hệ gần

PAM250

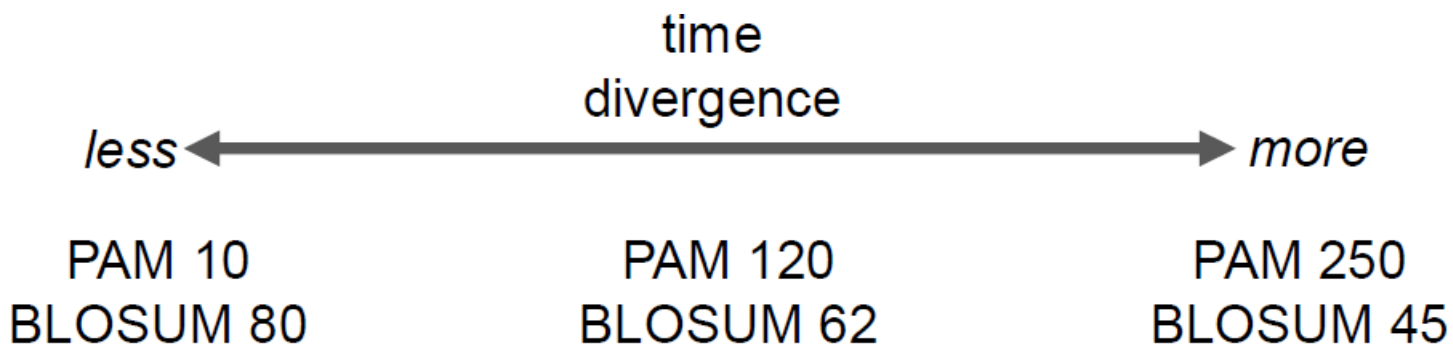
	A	R	N	D	C
A	2				
R	-2	6			
N	0	0	2		
D	0	-1	2	4	
C	-2	-4	-4	-5	12

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

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Ma trận thay thế PAM và BLOSUM



- PAM number:
  - tương ứng với thời gian tiến hóa
  - Giá trị càng lớn thì càng xa thời điểm từ gốc tổ tiên chung
- BLOSUM number:
  - Sự giống nhau về trình tự
  - Giá trị càng lớn thì các trình tự càng giống nhau

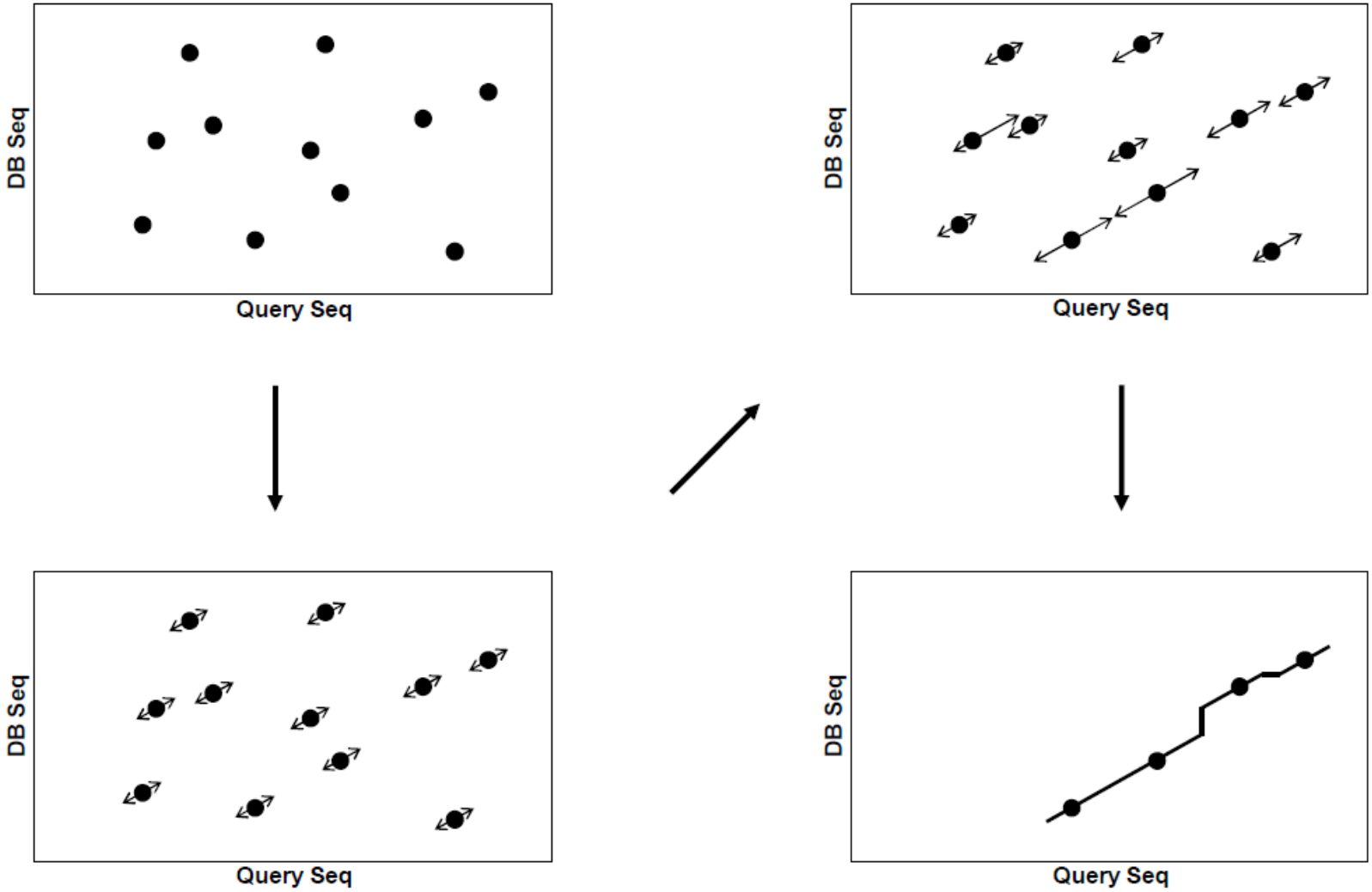
Phương pháp tìm kiếm các trình tự tương đồng trong CSDL

- Trong CSDL NCBI nr/nt có hơn 337 tỉ trình tự nucleotides
- Các phương pháp tìm kiếm vét cạn sẽ tổn thời gian
- => Các phương pháp heuristics: chỉ tìm một phần nhỏ có khả năng cho score lớn trong không gian tìm kiếm
- Phương pháp heuristics phổ biến nhất: BLAST

Basic Local Alignment Search Tool (BLAST)

- Mục đích: tìm ra các High Scoring Segment Pairs (HSPs) giữa trình tự truy vấn với các trình tự trong CSDL
- Ý tưởng:
  - True matches có khá năng chứa các đoạn ngắn tương tự nhau
  - Từ các đoạn ngắn này được sử dụng để tạo ra đoạn tương đồng dài
  - Các đoạn ngắn này có thể index trước trong CSDL

Ý tưởng của BLAST



Bước 1. Tạo danh sách các synonym

```
>query
ASGDAAGVSEQTPKLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTENCQDAPACNELRELAE
ASG
 SGD
  GDA
   DAA
    AAG
     . . .
```

words

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

BLOSUM62 matrix

Threshold Word Score = 13

			H	E	A	score
word						
H	E	A	8	5	4	17
synonyms						
H	D	A	8	2	4	14
H	Q	A	8	2	4	14
H	K	A	8	1	4	13
H	E	S	8	5	1	14
H	E	C	8	5	0	13
H	E	T	8	5	0	13
H	E	G	8	5	0	13
H	E	V	8	5	0	13
unacceptable synonym						
Y	E	A	2	5	4	11

Bước 2. Tìm các trình tự tiềm năng trong CSDL

Expanded Word List

HEA
HDA
HQA
HKA
HES
HEC
HET
HEG
HEV
...

Sequence Database

```
>seq1
GTKCCEFQAKLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQVGFGIATAVASGAA
>seq2
FQAKLFGIATAVASEHLLGAGTKCCEAPMCQFLAGKPEQVASVRIILEHGTRETLN
>seq3
QAKLFGFQAKLFGIAHECASEVAGKPVASRETLEAPMCQGKPVASVRIILETRETLNEQIILEHG
>seq4
VASVRGAGTKCCEAPMCQFLAGKPVASVRIILETRETLNEQIILEHGHLLFLAGKPVTL
>seq5
HGHLLFLAGKPVASRETLEAGAGTKCCEHGHLLFLAGKPVASVRIILETAGKPVASVRI
>seq6
ILETRETLNGKPVASRETLEAPMCQEQIIHEVVAGKPVASRETLEAPMCQG
>seq7
LETHGHLLFLAGPMCQFLVRIIRIILETAGKPVASVRIKPVASVHGHLLFLAGKPETLEARET
```

Bước 3. Tìm HSP

```
>query
ASGDAAGVSEQTPKLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTENCQDAPACNELRELAE
>seq1
GTKCCEFQAKLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQVGFGIATAVASGAA
```

Align at seed word.

```
ASGDAAGVSEQTPKLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTENCQDAPACNELRELAE
                                   H+A
    GTKCCEFQAKLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQVGFGIATAVASGAA
```

Extend sequence to identify HSP = High-scoring Segment Pair

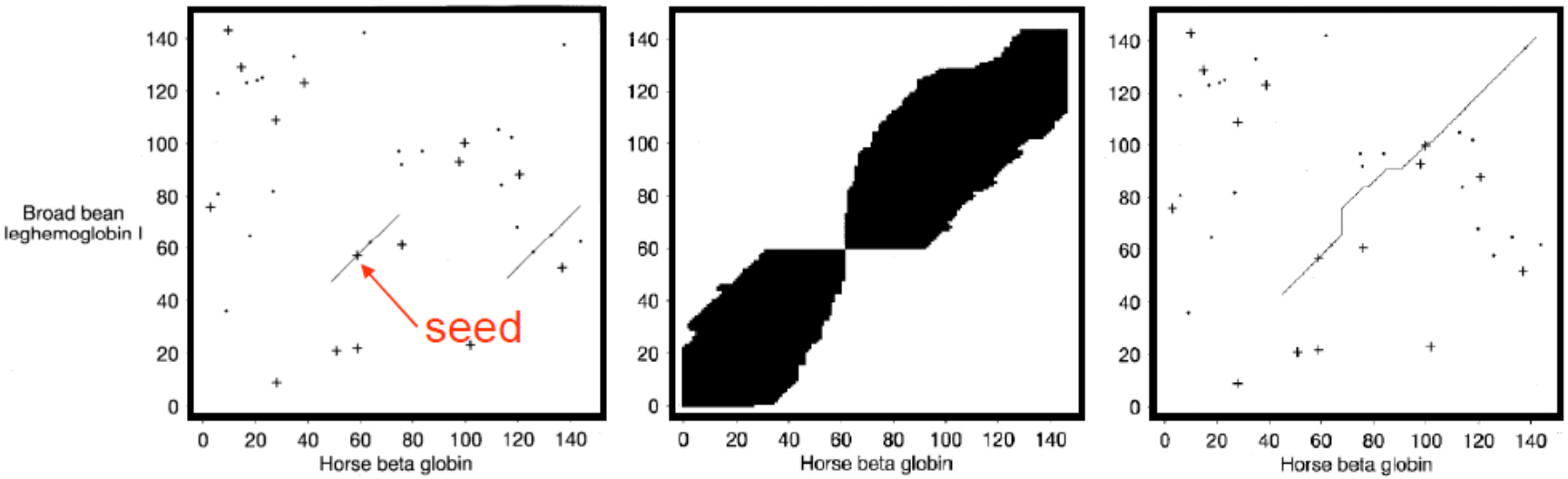
```
ASGDAAGVSEQTPKLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTENCQDAPACNELRELAE
             KL Q+LA KPEHP+ R+ L+A H+A +I++   L +
    GTKCCEFQAKLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQVGFGIATAVASGAA
```

Stop extension of HSP when quality of the alignment reaches a threshold value. Calculate significance.

```
KLAQYLADKPEHPLNRQRLDAKHEAFKIVVLQALTE
KL Q+LA KPEHP+ R+ L+A H+A +I++   L +
KLCQFLAGKPEHPMTRETLNASHQAVRIILEHLLEQ

Score = 42.0 bits (97), Expect = 0.004
Identities = 17/36 (47%), Positives = 26/36 (72%)
```

BLAST algorithm – gaps



seed

```
Leghemoglobin  43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------  90
                  F  L +   V+ +PK+ AH +KV          L + GE V  LD   G+
Beta globin    45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHHLDNLKGTFAALSE  90

Leghemoglobin  91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
                  +H  K  +DP +F ++   L+  +    G  ++ EL A+++    G+A A+
Beta globin    91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

Altschul, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25: 3389-3402

BLAST - programs

Program	Query	Database	Alignment	# Searches	Uses
blastn	DNA	DNA	DNA	1	find homologous DNA sequences
tblastx	DNA	DNA	protein	36	find homologous proteins from unannotated query and db sequences
blastx	DNA	protein	protein	6	identify proteins in query DNA sequence
tblastn	protein	DNA	protein	6	find homologous proteins in unannotated DNA DB
blastp	protein	protein	protein	1	find homologous proteins

```
                    5'-GTCACGTTACCGGTGGCCGAACAGGCCCGTCATGAAGT-3'
1st reading frame → V  T  L  P  V  A  E  Q  A  R  H  E  V
 2nd reading frame → S  R  Y  R  W  P  N  R  P  V  M  K  X
  3rd reading frame → H  V  T  G  G  R  T  G  P  S  *  S

                    5'-GTCACGTTACCGGTGGCCGAACAGGCCCGTCATGAAGT-3'
                    3'-CAGTGCAATGGCCACCGGCTTGTCCGGGCAGTACTTCA-5'
                       T  V  N  G  T  A  S  C  A  R  *  S  T ← 4th reading frame
                      X  *  T  V  P  P  R  V  P  G  D  H  L ← 5th reading frame
                        D  R  *  R  H  G  F  L  G  T  M  F ← 6th reading frame
```

BLAST – databases

Protein Databases

nr	Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF
swissprot	Last major release of the SWISS-PROT protein sequence database
pat	Proteins from the Patent division of GenBank.
month	All new or revised GenBank CDS translations + PDB + SwissProt + PIR + PRF released in the last 30 days.
pdb	Sequences derived from the 3-dimensional structure records from the Protein Data Bank

Nucleotide Databases

nr/nt	All GenBank + EMBL + DDBJ + PDB + RefSeq sequences (but no EST, dbSTS, GSS, WGS,TSA or phase 0, 1 or 2 HTGS sequences).
est	Database of GenBank + EMBL + DDBJ sequences from EST division
refseq_rna	NCBI transcript reference sequences
refseq_representative_genomes	Reference and representative genomes selected from the NCBI Refseq Genomes database
gss	Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
htgs	Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2. Finished, phase 3 HTG sequences are in nr.
pat	Nucleotides from the Patent division of GenBank.
pdb	Sequences derived from the 3-dimensional structure records from Protein Data Bank.
tsa	Transcriptome Shotgun Assembly (TSA) database is an archive of computationally assembled mRNA sequences
sra	Search for sequences associated with a particular SRA (sequence read archive) accession, scientific name, or taxonomic identifier
dbsts	Database of Sequence Tag Site entries from the STS division of GenBank + EMBL + DDBJ.
refseq_genomes	NCBI Refseq genomes across all taxonomy groups. Contains only the top-level sequences, i.e. chromosomal sequences where available (but not the contigs used to assemble them)
wgs	Assemblies of Whole Genome Shotgun sequences

PSI-BLAST – position-specific iterated-BLAST

Motif or profile search methods are often more sensitive than pairwise comparisons at detecting distant relationships.

Most useful for finding protein families.

Process

- Create a multiple sequence alignment from BLAST output
- Use the MSA to automatically create a position-specific scoring matrix (PSSM)
  - generated by identifying conserved columns in MSA
- Use PSSM to score BLAST search
- Iterate

PSI-BLAST

```
730496    66 FTVDENGQMSATAKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDH 125
200679    63 FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 122
206589    34 FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 93
2136812    2         MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 53
132408    65 FKIEDNGKTTATAKGRVRILDKLELCANMVGTFIETNDPAKYRMKYHGALAILERGLDDH 124
267584    44 FSVDESGKVTATAHGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAASYLQTGNDDH 103
267885    44 FSVDGSGKVTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKERYWGAAAYLQSGNDDH 103
8777608   63 FTIHEDGAMTATAKGRVIILNNWEMCADMMATFETTPDPAKFRMRYWGAASYLQTGNDDH 122
6687453   60 FKVEEDGTMTATAIGRVIILNNWEMCANMFGTFEDTEDPAKFKMKYWGAAAYLQTGYDDH 119
10697027  81 FKVQEDGTMTATATGRVIILNNWEMCANMFGTFEDTEEPARFKMKYWGAAAYLQTGYDDH 140
13645517  1                                MVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 32
13925316  36 FSVDGSGKMTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 97
131649    65 YTVEEDGTMTASSKGRVKLFGFWVICADMAAQYTDPTTPAKMYMTYQGLASYLSSGGDNY 126
```

R,I,K C D,E,T K,R,T N,L,Y,G

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Pos	aa	Amino Acid																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	-2	-3	-2	-1	-2	-3	-2	**1**	**2**	-2	**6**	0	-3	-2	-1	-2	-1	**1**
2	K	-1	**1**	0	**1**	-4	**2**	**4**	-2	0	-3	-3	**3**	-2	-4	-1	0	-1	-3	-2	-3
3	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	0	-4	-3	-3	**12**	**2**	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4	-4	**3**	**1**	-3	**1**	-1	-3	-2	0	-3	-1	**4**
5	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	0	-4	-3	-3	**12**	**2**	-3
6	A	**5**	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	**1**	0	-3	-2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

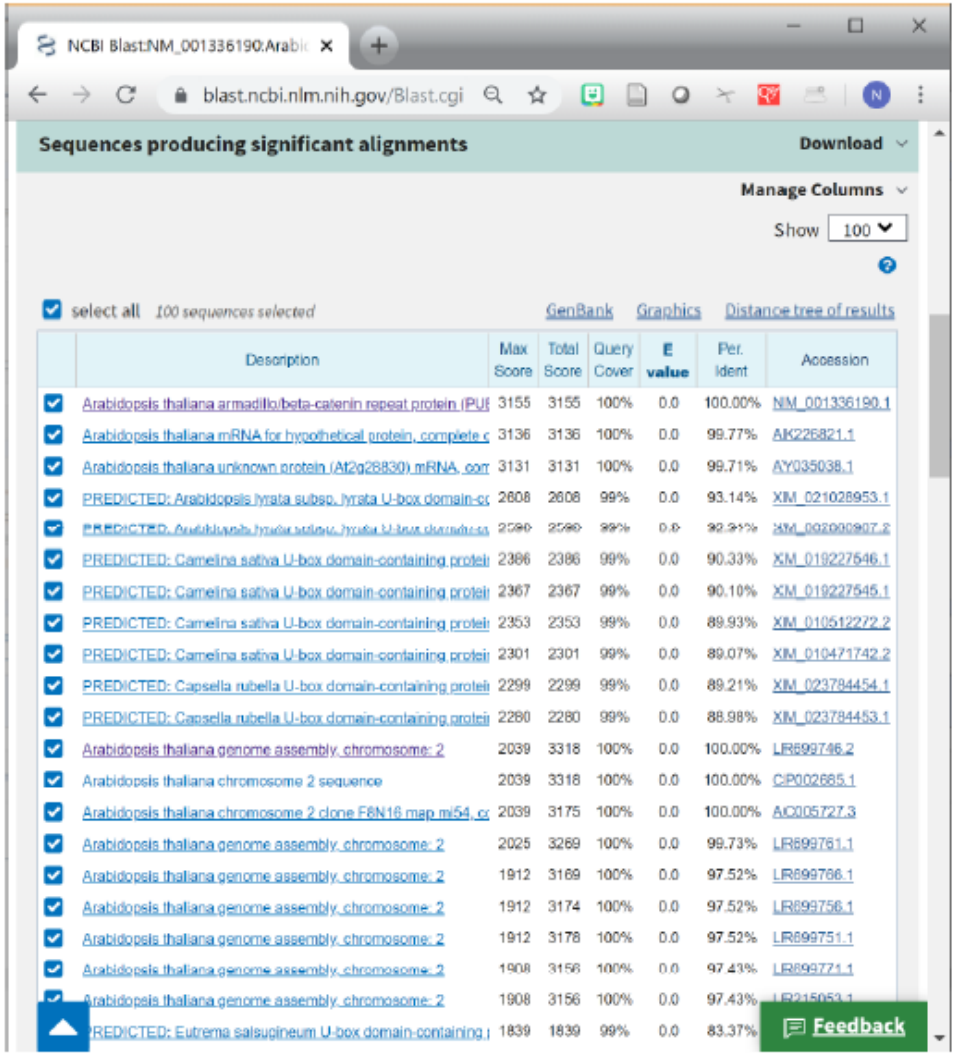
Pevsner 2003 Bioinformatics and Functional Genomics. Wiley-Liss

Evaluation of BLAST results

Is a DB sequence homologous to the query?

- significant expect values
- reciprocal best hit
- similar sizes
- common motifs
- reasonable multiple sequence alignment
- similar 3D structures

Is one DB hit better than another?

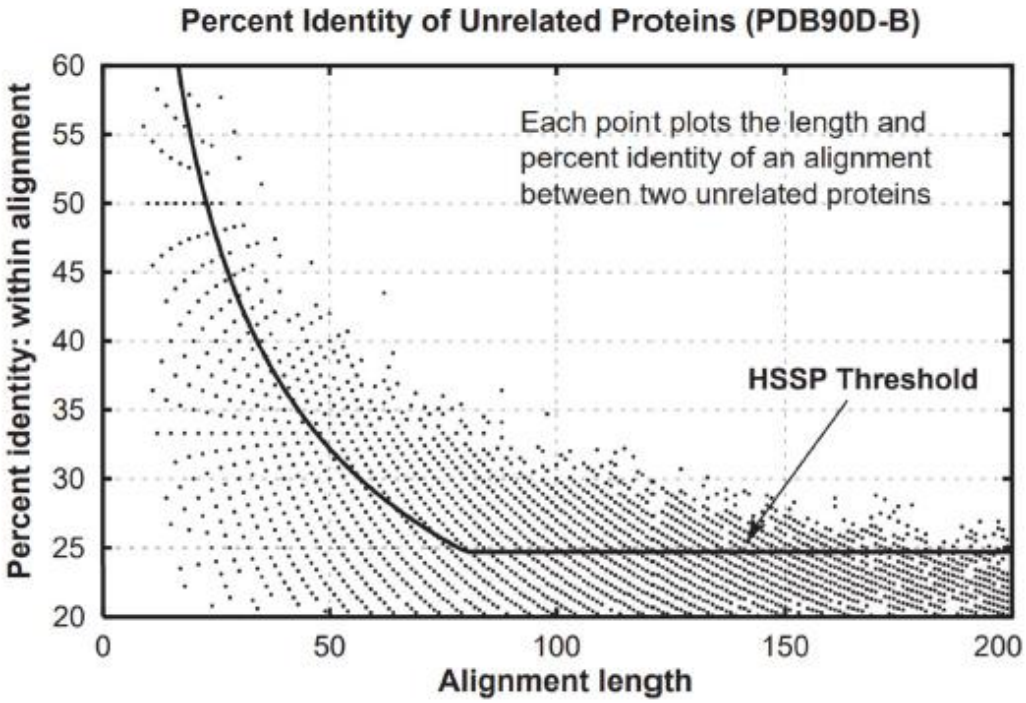


Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Arabidopsis thaliana armadillo/beta-catenin repeat protein (PUI	3155	3155	100%	0.0	100.00%	NM_001336190.1
Arabidopsis thaliana mRNA for hypothetical protein, complete c	3136	3136	100%	0.0	99.77%	AK226821.1
Arabidopsis thaliana unknown protein (At2g28830) mRNA, com	3131	3131	100%	0.0	99.71%	AY035038.1
PREDICTED: Arabidopsis lyrata subsp. lyrata U-box domain-co	2608	2608	99%	0.0	93.14%	XM_021028953.1
PREDICTED: Arabidopsis lyrata subsp. lyrata U-box domain-co	2590	2590	99%	0.0	92.91%	XM_002890907.2
PREDICTED: Camelina sativa U-box domain-containing protei	2386	2386	99%	0.0	90.33%	XM_019227546.1
PREDICTED: Camelina sativa U-box domain-containing protei	2367	2367	99%	0.0	90.10%	XM_019227545.1
PREDICTED: Camelina sativa U-box domain-containing protei	2353	2353	99%	0.0	89.93%	XM_010512272.2
PREDICTED: Camelina sativa U-box domain-containing protei	2301	2301	99%	0.0	89.07%	XM_010471742.2
PREDICTED: Capsella rubella U-box domain-containing protei	2299	2299	99%	0.0	89.21%	XM_023784454.1
PREDICTED: Capsella rubella U-box domain-containing protei	2280	2280	99%	0.0	88.98%	XM_023784453.1
Arabidopsis thaliana genome assembly, chromosome: 2	2039	3318	100%	0.0	100.00%	LR699746.2
Arabidopsis thaliana chromosome 2 sequence	2039	3318	100%	0.0	100.00%	CP002685.1
Arabidopsis thaliana chromosome 2 clone F8N16 map mi54, co	2039	3175	100%	0.0	100.00%	AC005727.3
Arabidopsis thaliana genome assembly, chromosome: 2	2025	3269	100%	0.0	99.73%	LR699761.1
Arabidopsis thaliana genome assembly, chromosome: 2	1912	3169	100%	0.0	97.52%	LR699766.1
Arabidopsis thaliana genome assembly, chromosome: 2	1912	3174	100%	0.0	97.52%	LR699756.1
Arabidopsis thaliana genome assembly, chromosome: 2	1912	3178	100%	0.0	97.52%	LR699751.1
Arabidopsis thaliana genome assembly, chromosome: 2	1908	3156	100%	0.0	97.43%	LR699771.1
Arabidopsis thaliana genome assembly, chromosome: 2	1908	3156	100%	0.0	97.43%	LR215053.1
PREDICTED: Eutrema salsugineum U-box domain-containing p	1839	1839	99%	0.0	83.37%	

Statistical evaluation – sequence identity?

Why not use sequence identity?

- distribution not well understood
- difficulty with shared domains that do not stretch over length of sequence
- false positive rate
- ignores gaps and conservative vs. radical substitutions



Brenner et al. 1998 PNAS 95:6073

Statistical evaluation – bit score

BLAST reports two bit scores, *S* and *R*

a, b defaults are 1, -2 for Blastn; a slightly different formula and substitution matrices are used for protein bit scores

Raw bit scores (*R*)

$$R = aI + bX - cO - dG$$

I = # identities in the alignment

X = # mismatched residues

O = # gaps

G = # of '-' (length of gap)

a = reward for each identity

b = 'reward' for each mismatch

c = gap opening penalty

d = penalty for each '-'

Can be adjusted manually in Blast

Statistical evaluation – bit score

Normalized bit scores (S)

$$S = (\lambda R - \ln K) / (\ln 2)$$

λ and K are normalizing parameters

- λ is a scale factor which converts pairwise match scores to probabilities
- K is a proportionality constant to correct for the number of sequence comparisons

Makes bits scores (and E-values) independent of the scoring system

Available from Blast Search Summary

Statistical evaluation – E value

Expect (E) values – best measure of significance

Converts a bit score into a probability

Depend upon

- Bit Score (S)
- Effective length of query (m)
- Effective length of database (n)

$$E = mn2^{-S}$$

Probability of finding a database match as good as or better than your query by chance.

How Good is My Hit?

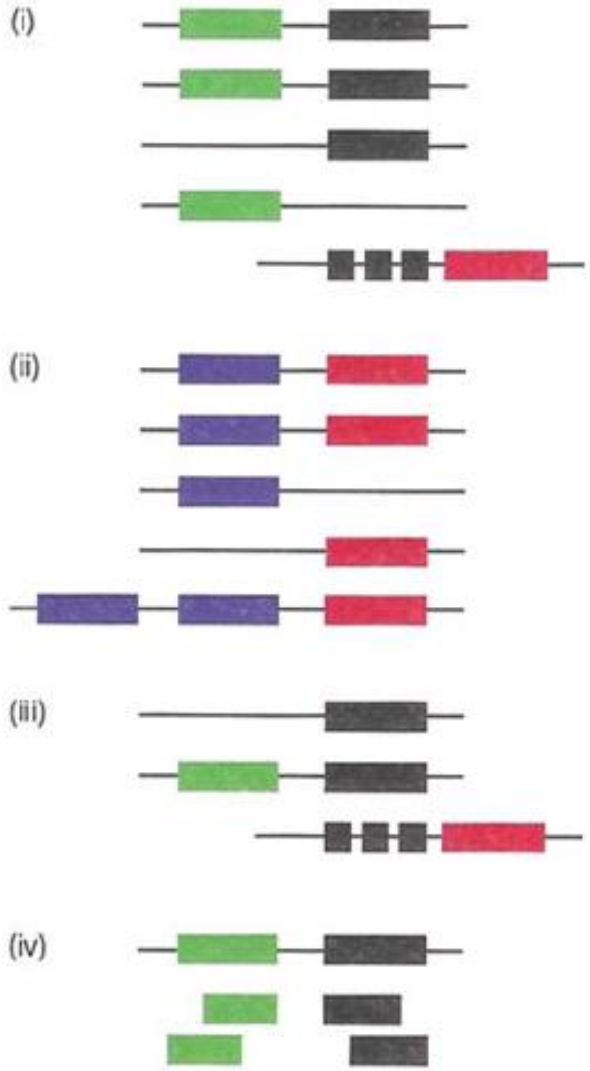
Use identity? No!

Use bit score: better

Use E value: best

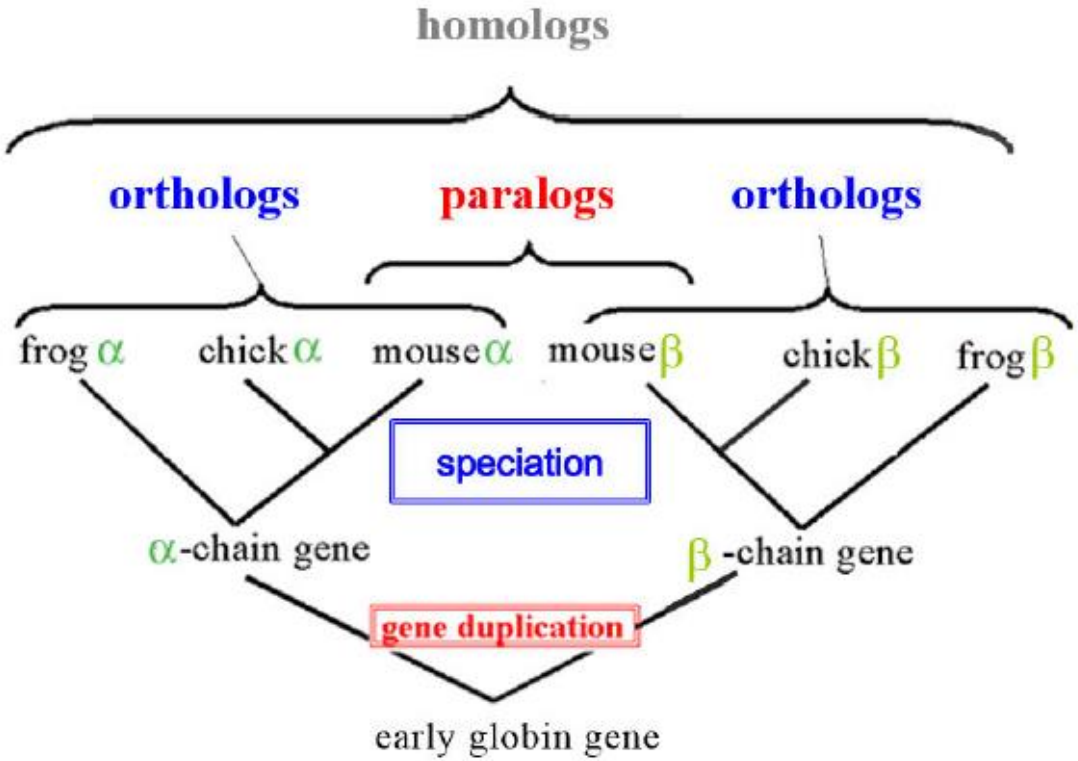
- The E value is a probability value that is based on the number of different alignments with scores at least as good as that observed, which are expected to occur simply by chance.
- The lower the E value, the more significant the score. This is by far the best metric to use since results of different searches in the same database can be readily compared.
- Note that E value is dependent on the size of the database (n) and the length of the query sequence (m). *The same sequence searched on different databases containing identical hit sequences would result in different E values being reported for those sequences.*

What is a good E value?



	Amino acid alignment	Sequence number	Typical range of P/E value[7]
(i)		1 (query)	—
		2	$<10^{-20}$
		3	$10^{-8} - 10^{-20}$
		4	$10^{-8} - 10^{-20}$
		5	$10^{-6} - 10^{-8}$
(ii)		6 (query)	—
		7	$<10^{-20}$
		8	$10^{-8} - 10^{-20}$
		9	$10^{-8} - 10^{-20}$
		10	$<10^{-20}$
(iii)		3 (query)	—
		1,2	$10^{-8} - 10^{-20}$
		5	$10^{-6} - 10^{-8}$
(iv)		1 (query)	—
		EST hits	$<10^{-4}$

Orthology and Paralogy

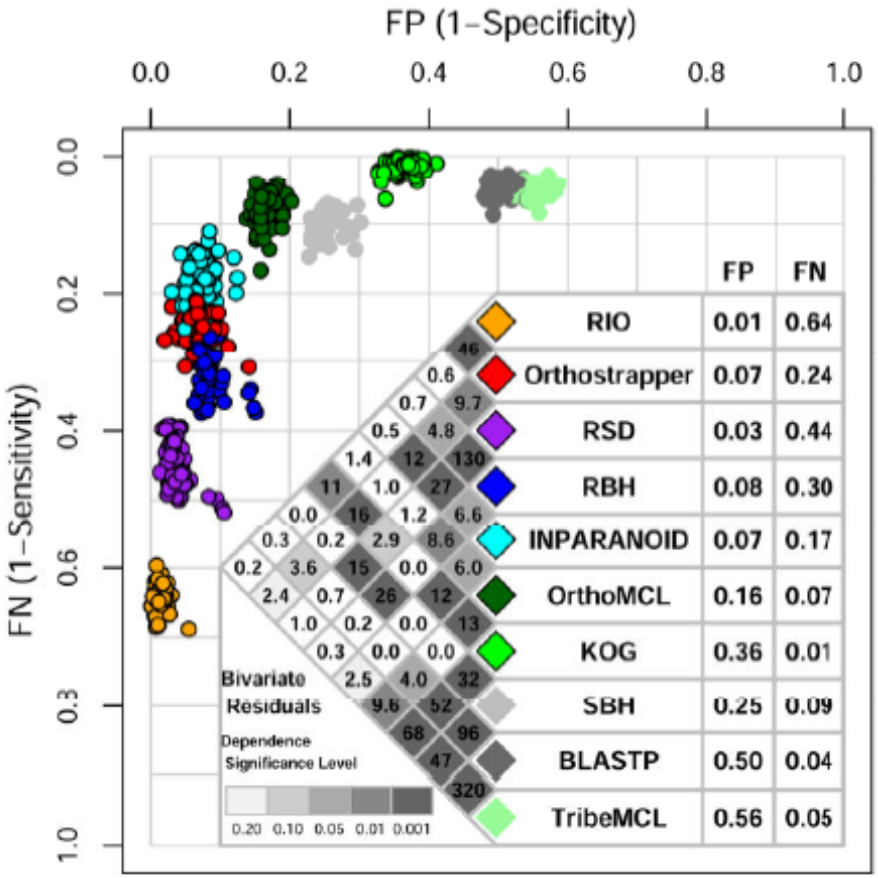


- Orthology can be used to identify conserved residues within genes and proteins
- In addition, comparative genomic methods can be applied to intron sequences and promoters to identify parts of these that are conserved and hence potentially functionally important

Methods for determining orthology in genomic sequences

- TBLASTX or BLASTP – take reference genome and blast against other genomes, and take region (gene) with best e-value (above a threshold) as orthologous region. Problem: what if blasting in other direction identifies a match in reference genome that is better? Which is the ortholog?
- Reciprocal Best Hit (RBH) method – addresses the above issue but can get confounded by rampant domain swapping that has occurred, esp. in eukaryotic genomes → lots of false negatives
- Phlyogenetic-based methods such as RIO, Orthostrapper and RSD
- BLASTP-based methods, such as InParanoid, OrthoMCL, KOG: these use BLASTP followed by Markov or other Clustering methods

Overview of the methods: which is best?



from Chen et al., 2007, PLoS One 2(4): e383

Other tools are available, e.g. OrthoFinder2

(Emms & Kelly, 2019, https://dx.doi.org/10.1186/s13059-019-1832-y)

→ what are FP and FN rates for any tool you might want to use?

Ortholog databases

Clusters of Orthologous Groups (COG) and euKaryotic Orthologous (KOG) Groups: http://www.ncbi.nlm.nih.gov/COG/ *several species, updated 2020

HieranoiDB: http://hieranoidb.sbc.su.se/ *66 species, slightly older

Kaduk M, Riegler C, Lemp O, Sonnhammer EL. HieranoiDB: a database of orthologs inferred by Hieranoid. Nucleic Acids Res. 2017, 45(Database issue), D687-D690. PMID: 27742821.

OrthoMCL DB: http://www.orthomcl.org/ *many species, Nov. 2020 release

Feng Chen, Aaron J. Mackey, Christian J. Stoeckert, Jr and David S. Roos. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Research 2006 34(Database Issue):D363-D368

InParanoid DB: http://inparanoid.sbc.su.se/cgi-bin/index.cgi *273 species, from 2013

Remm M, Storm CEV and Sonnhammer ELL (2001). Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. JMB, 314:1041-1052.

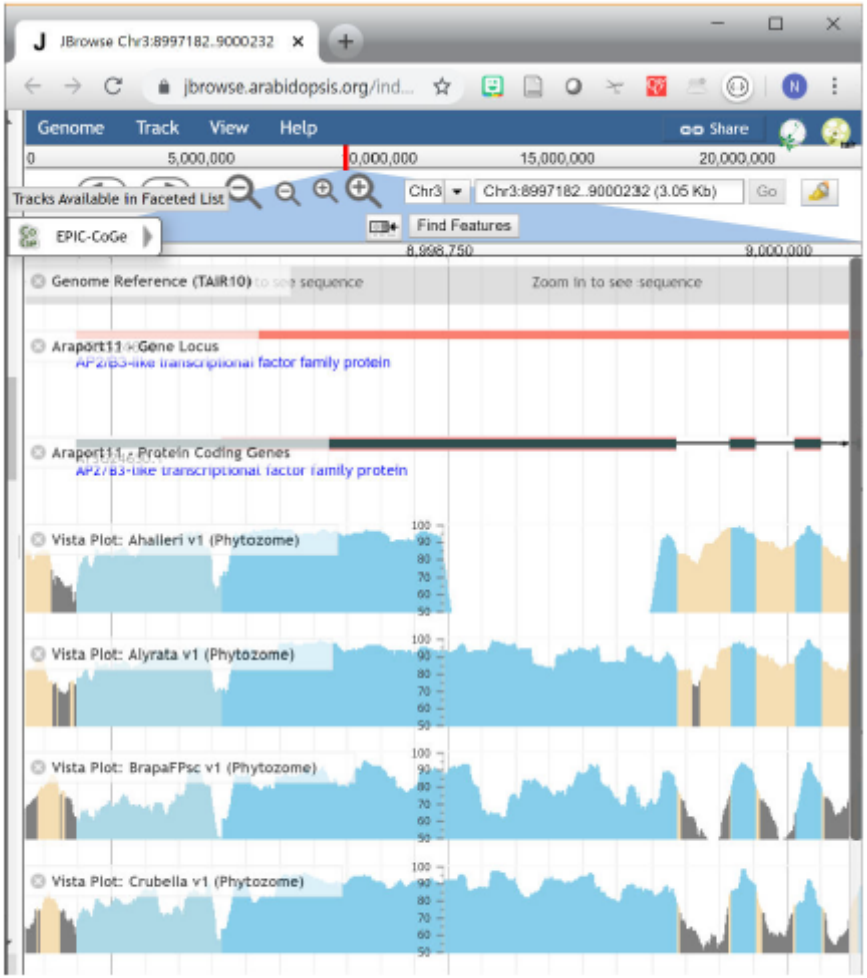
CoGe: http://genomeevolution.org/ 50,000+ genomes, up-to-date; syntenty tools!

Lyons E ~ Lisch D (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids, Plant Phys 148, pp. 1772–1781.

You may find others → how up-to-date are these, genome versions, etc.?

Tools for comparative genomics & genome browsing

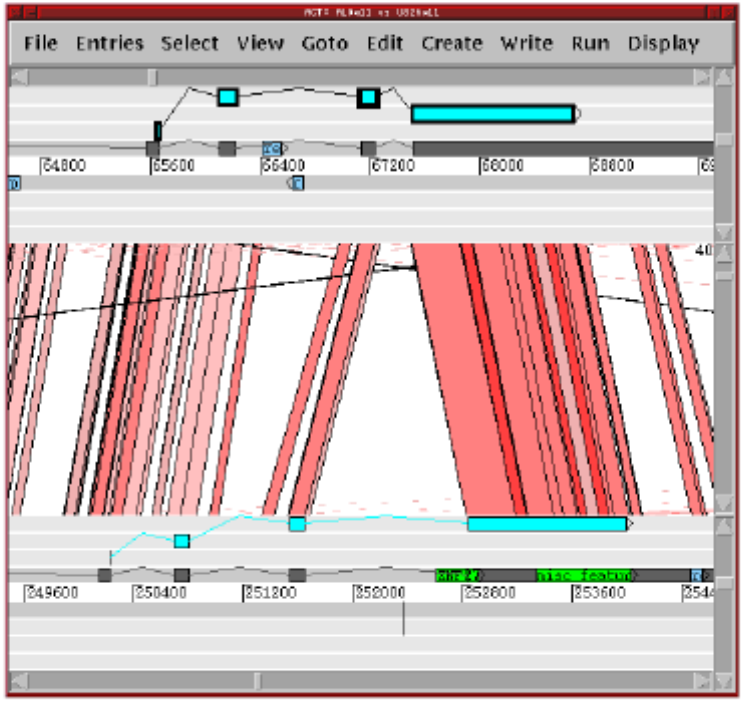
GBrowse/JBrowse is standard for many model organisms



ACT (Artemis Comparison Tool) standalone tool that allows cross-genome comparisons

http://www.sanger.ac.uk/science/tools/artemis

allows rearrangements and syntenic blocks to be easily visualized



Genome comparisons and synteny

- Synteny is the preservation of gene order on chromosomes of related species

- During evolution, genomic rearrangements can separate two loci
  → result is a loss of synteny between them

- Translocations can also join two previously separate pieces of chromosomes (rare event)
  → results in a gain of synteny between loci

- Synteny can be useful in the case of many-to-many or one-to-many ortholog mappings, for determining the "true" ortholog, and also identifying translocations/ inversions – these show up as blocks which cross other blocks, and as "X" shaped figures in e.g. ACT

Potential translocations/inversions