Database searches

1

#### DNA and protein databases

• EMBL/GenBank/DDBJ database of nucleic acids



## DNA and protein databases

• EMBL/GenBank/DDBJ database of nucleic acids (cntd)



#### DNA and protein databases

• SWISS-PROT & TrEMBL database of proteins



Number of entries in UniProtKB/TrEMBL

## Searching databases: motivation

- Determine orthologs and paralogs for a protein of interest, assign putative function
  - A new bacterial genome is sequenced, how many genes have related genes in other species
- Determine if a genome contains specific types of proteins
- Determine the identity of a DNA or protein sequence
  - What is the identity of a clinical pathogen?
- Determine if particular variant has been described before
  - Many pathogens, especially viruses, mutate rapidly. We should like to know if we have a new strain.

## BLAST: Basic Local Alignment Search Tool

- Performs many alignments at once
- Heuristic algorithms are used instead of DP. Why?
  - Size of SWISS-PROT + TrEMBL (Rel. 9.5):
    3.9M entries or 1,276M residues.
  - Exact algorithms are O(NM) fast.
- Heuristic methods can look at a small fraction of the searching space that will include all (or most) of the high scoring pairs.
- Web interface and standalone program

## BLASTP algorithm

- Idea: search space is reduced by looking only at small exact (or near exact matches)--words
- List compile the list of words using the query
- Scan –scan the reference for matches
  - HSP-high scoring segment pairs
  - Location is not stored at this point
- Extend
  - Ungapped extension-simple
  - Gapped extension using dynamic programing
- Report hits above threshold

Phase 1: Setup: compile a list of words (w=3) above threshold T

• Query sequence: human beta globin NP\_000509.1 (includes ...VTALWGKVNVD...). This sequence is read; low complexity or other filtering is applied; a "lookup" table is built.

• Words derived from query sequence (HBB): VTA TAL ALW LWG WGK GKV KVN VNV NVD

Generate a list of words matching query		LWG	4+11+6=21
(beth chows and below T) Consider THG		IWG	2+11+6=19
in the query and the secret (derived from a		MWG	2+11+6=19
In the query and the scores (derived normal		VWG	1+11+6=18
BLOSOW62 matrix) for various words.	examples of	FWG	0+11+6=17
· Concrete similar lists of words epopping	words >=	AWG	0+11+6=17
Generate similar lists of words spanning	threshold 12	LWS	4+11+0=15
the query (e.g. words for wGw, GwG, wGR).		LWN	4+11+0=15
		LWA	4+11+0=15
throchold		LYG	4+ 2+6=12
threshold		LFG	4+ 1+6=11
	examples of	FWS	0+11+0=11
	words below	AWS	-1+11+0=10
	threshold	CWS	-1+11+0=10
		TWC	2+11-3=10

Phas • Sel • Sca • Cre • Pel	se 2: Scanning and extensions lect all the words above threshold T (LWG, IWG, MWG, VWG, FWG, AWG, LWS, LWN, LWA, LYG) an the database for entries ("hits") that match the compiled list eate a hash table index with the locations of all the hits for each word form gap free extensions
• Per	rform gapped extensions
	LTPERKSAVTALWGKVNVDEVGGEALGRLLVVYPWTORFFESFGDLSTPDAVMGNPKV HBB
	L+P +K+ V A WGKV + E G EAL R+ + +P T+ +F F D G+ +V
	LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQV HBA
	extension extension
	first phases of search
	"hits" alpha globin,
	triggers extension

#### Phase 3: Traceback

- Calculate locations of insertions, deletions, and matches (for alignments saved in Phase 2)
- Apply composition-based statistics (for BLASTP, TBLASTN)
- Generate gapped alignment

## Compositional adjustment

- Recall the log odds score
- Background frequencies should be the the marginal frequencies of q<sub>ij</sub>
- Compositional adjustment
  - Use empirical  $p_i p_j$
  - Adjust  $\underline{q}_{ij}$  accordingly

$$s_{ij} = \frac{1}{\lambda} \ln \left( \frac{q_{ij}}{p_i p_j} \right),$$

$$p_i = \sum_j q_{ij}; \quad p'_j = \sum_i q_{ij};$$

Yu et al. The compositional adjustment of amino acid substitution matrices

#### Compositional adjustment

	Organisms compared	No. of	Mean BLOSUM-62 bit score*	Background	Median change in bit score* with respect to BLOSUM-62		Cases improved	Cases (%) with statistical significance
Sequence pairs		pairs		specified	Absolute	Relative (%)	(%)	a factor >10 <sup>†</sup>
Related	C. tetani and M. tuberculosis	40	68.3	Organism <b>Sequence</b> ‡	+1.6 + <b>2.3</b>	+2.7 +3.3	58 <b>85</b>	20/8 <b>38/3</b>
	B. subtilis and L. lactis	37	59.8	Organism <b>Sequence</b> ‡	+1.1 + <b>2.1</b>	+1.8 +3.6	84 <b>95</b>	16/3 <b>11/3</b>
	M. tuberculosis and S. coelicolor	34	58.6	Organism <b>Sequence</b> <sup>‡</sup>	+1.4 + <b>2.7</b>	+2.6 + <b>4.1</b>	76 <b>100</b>	24/3 <b>32/0</b>
Unrelated (negative control)	C. tetani and M. tuberculosis	1,560	16.7	Organism Sequence <sup>‡</sup>	-0.02 -0.05	-0.1 -0.3	49 47	0.4/0.1 0.6/0.4
	B. subtilis and L. lactis	1,332	15.7	Organism Sequence <sup>‡</sup>	+0.00 +0.04	+0.0 +0.3	50 52	0.0/0.0 0.2/0.4
	M. tuberculosis and S. coelicolor	1,122	16.4	Organism Sequence <sup>‡</sup>	+0.05 +0.06	+0.3 +0.4	53 53	0.0/0.1 0.6/0.2
Structural	Various	32	50.4	Sequence <sup>‡</sup>	+1.3	+3.2	72	22/0

Yu et al. The compositional adjustment of amino acid substitution matrices

(b) Query: human insulin NP\_000198Program: blastpDatabase: *C. elegans* RefSeqOption: No compositional adjustment

> ref | NP\_501926.1 | UG INSulin related family member (ins-1) [Caenorhabditis elegans] Length=109 Score = 34.7 bits (78), Expect = 0.009 Identities = 30/100 (30%), Positives = 41/100 (41%), Gaps = 14/100 (14%) Query 11 LALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGG 70 LA+L L P P+ A + LCGS L L VC + +R A+ LAILLLSSPTPSDASIR--LCGSRLTTTLLAVCRNQLCTGLTAFKRSADQSY----- 66 Sbjct 17 Query 71 GPGAGSLQPLALEGSLQKRG-IVEQCCTSICSLYQLENYC 109 A + + LQKRG I +CC CS L+ +C Sbjct 67 ---APTTRDLFHIHHQQKRGGIATECCEKRCSFAYLKTFC 103

(c) Query: human insulin NP\_000198
Program: blastp
Database: *C. elegans* RefSeq
Option: conditional compositional score matrix adjustment

> ref | NP\_501926.1 | UG INSulin related family member (ins-1) [Caenorhabditis elegans] Length=109

```
Score = 33.5 bits (75), Expect = 0.020, Method: Compositional matrix adjust.
Identities = 27/100 (27%), Positives = 39/100 (39%), Gaps = 12/100 (12%)
```

Query	10	LLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELG LA+L L P P+ A + LCGS L L VC + +R A+	69
Sbjct	16	FLAILLLSSPTPSDASIRLCGSRLTTTLLAVCRNQLCTGLTAFKRSADQS	65
Query	70	GGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYC 109 P L + ++ GI +CC CS L+ +C	
Sbjct	66	YAPTTRDLFHIHHQQKRGGIATECCEKRCSFAYLKTFC 103	
		V	

# **BLAST flavors**

- BLASTN-requires exact word matched
  - Length 7-15 (11 default)
  - Requires two word pairs within some distance
    - Returns more hits but saves time in the extension phase



#### **BLAST** statistics

- Parameters to consider
  - Length of query longer queries will generate more matches
  - Database size
- Raw score
- Bit score –parameter normalized score comparable across searches
- E value-expected number of sequences
- P-value-probability of a chance alignment occurring with this score or better

## How to interpret a BLAST search: expect value

- It is important to assess the statistical significance of search results
- For global alignments, the statistics are poorly understood.
- For local alignments (including BLAST search results), the statistics are well understood.
- The scores follow an extreme value distributio (EVD)
  - Theoretically for ungapped alignments
  - Empirically for gapped alignments



## Calculating E

- $E = Kmn e^{-\lambda S}$
- This equation is derived from a description
- of the extreme value distribution
- S =the score
- m, n = the length of two sequences
- $\lambda$ , K = Karlin Altschul statistics

## E vs P

- Very small *E* values are very similar to *p* values.
- *E* values of about 1 to 10 are far easier to interpret
- than corresponding *p* values.

• <u>E</u>	<u>p</u>
• 10	0.99995460
• 5	0.99326205
• 2	0.86466472
• 1	0.63212056
• 0.1	0.09516258 (about 0.1)
• 0.05	0.04877058 (about 0.05)
• 0.001	0.00099950 (about 0.001)
• 0.0001	0.0001000

## Problems that BLAST can't solve

- Finding very distant homologs
  - Human myoglobin does not come up as a hit when searching with beta-globin even though they share the same structure
  - To little similarity
- Aligning large genomic segments
  - Align large chromosomal regions between mouse and human genome
  - Some regions have high similarity while others do not
- Aligning nextgen sequencing data
  - Millions of 100bp reads to 3 billion bp of human gnome time to run BLASTN –weeks!

## Distant homologs

- Position specific iterated blast PSI-BLAST
  - Perform initial search with BLASTP
  - Perform a multiple sequence alignment with results
  - Define a position-specific scoring matrix PSSM
    - Reference position if fixed
    - One dimension represents position along reference instead of amino acids
  - Use the PSSM to search the database
  - Repeat



#### PSI-BLAST globin family



## PSI-BLAST problems and pitfalls

- Iterative algorithm errors propagate
- Low entropy regions regions with biased a.a composition can corrupt the PSSM
- Iterations can be adjusted by hand to remove suspicious sequences from the alignment

## Genomic alignment BLASTZ

- PatternHunter
  - Nonconsecutive seed boost sensitivity
  - Need 11 matches spanning 18 nucleotides
  - At specific positions
  - 110100110010101111
  - 64 nucleotides with 70% identity
    - Probability of BLASTN match 0.3
    - Probability of PatternHunter match 0.47
- Other considerations
  - Removing lineage specific repeats
    - Gene duplications
    - Retrotransposons
  - Masking already aligned regions

lllllllll ATGGTGCATCT (example of a seed)(extended) ATTGTGCATCT (example of a mismatch)(not extended)

(a)

(b) 1101001100101111 ATGGTGCATCTGACTCCT (example of a seed)(extended) ATTGTGCATCTGACTCCT (example of an acceptable match)(extended)



## Genome wide alignments



- Alignments can be viewed in UCSC
- Download alignment files and extract regions of interest
- If a regions(gene) is missing from an organism doesn't mean it is not there
  - Incomplete alignment
  - Incomplete genome assembly
  - Use BLAST!



## Short read alignment

- Bowtie -ultrafast, memoryefficient short read aligner
- Basic strategy
  - Index the genome
  - Use Burrows-Wheeler Transform BWT
  - Human genome fits entirely into RAM



## Why Burrows-Wheeler?

#### BWT very compact:

Approximately ½ byte per base

As large as the original text, plus a few "extras"

Can fit onto a standard computer with 2GB of memory

#### Linear-time search algorithm

proportional to length of query for exact matches



## Burrows-Wheeler Transform (BWT)

- Generate all circular permutations
- Sort by first letter
- Lest column is the BWT
- Everything else is discarded
- First column can be recovered from BWT
  - It has all the same characters sorted



Burrows-Wheeler Matrix (BWM)

#### Exact match

#### BWT(agcagcagact) = tgcc\$ggaaaac

gca gca gca gca \$agcagcagact **\$**agcagcagact \$agcagcagact \$agcagcagact act\$agcagcag act\$agcagcag act\$agcagcag act\$agcagcag agact\$agcagc agact\$agcagc agact\$agcagc agact\$agcagc agcagact\$agc agcagact\$agc agcagact\$agc agcagact\$agc agcagcagact\$ agcagcagact\$ agcagcagact\$ agcagcagact\$ cagact\$agcag cagact\$agcag cagact\$agcag cagact\$agcag cagcagact\$ag cagcagact\$ag cagcagact\$ag cagcagact\$ag ct\$agcagcaga ct\$agcagcaga ct\$agcagcaga ct\$agcagcaga gact\$agcagca gact\$agcagca gact\$agcagca gact\$agcagca gcagact\$agca gcagact\$agca gcagact\$agca gcagact\$agca gcagcagact\$a gcagcagact\$a gcagcagact\$a gcagcagact\$a t\$agcagcagac t\$agcagcagac t\$agcagcagac t\$agcagcagac

Search for pattern: gca

## Inexact matching

- If match cannot be extended try mismatches
- A->C-><mark>G</mark> X
- A->T-><mark>G</mark> X
- A->G->G 🗸

## Multiple Sequence Alignment

- Goal: Given several sequences bring the greatest number of similar characters into the same column of the alignment
- Why?
- Correspondence. Find out which parts "do the same thing"
  - Similar genes are conserved across widely divergent species, often performing similar functions
- Structure prediction
  - Use knowledge of structure of one or more members of a protein MSA to predict structure of other members
  - Structure is more conserved than sequence
  - Predict if mutations are deleterious by looking at cross species conservation
- Create "profiles" for protein families
  - Allow us to search for other members of the family– PSI-BLAST
- MSA is the starting point for evolutionary analysis

VTISCTGSSSNIGAG-NHVKWYQQLPG VTISCTGTSSNIGS—ITVNWYQQLPG LRLSCSSSGFIFSS—YAMYWVRQAPG LSLTCTVSGTSFDD—YYSTWVRQPPG PEVTCVVVDVSHEDPQVKFNWYVDG— ATLVCLISDFYPGA—VTVAWKADS— ATLVCLISDFYPGA—VTVAWKADS— XTLVCLISDFYPGA—VTVAWKADS—

## Multiple alignment problem

- Ribosome: an RNA/protein complex rpS14: a ribosomal protein in yeast
- Goal: Determine residues responsible for binding rpS14 to ribosomal RNA
- Known:
  - Sequence of rpS14
  - Structure of homolog in • bacteria
  - Sequences in many species •
- Find the MSA
- Find conserved residues
- Use structure to check for binding function

т.	thermophilus	МАК	KPSKKKVKRQVASGR	AYIHASYNNTIVTIT	PDGNPITWSSGGVI	GYKGSR-KGTPYANQ
л.	aeolicus	M	AKKKKKQKRQVTKAI	VHIHTTFNNTIVNVT	DTQGNTIAWASGGTV	GFKGTR-KSTFYAAQ
Ρ.	aeruginosa	MAKPA	ARPRKKVKKTVVDGI	AHIHASFNNTIVTIT	DRQGNALSWATSGGS	GFRGSR-KSTPFAAQ
Ε.	coli	MAKAP	IRARKRVRKQVSDGV	AHIHASENNTIVTIT	DRQGNALGWATAGGS	GFRGSR-KSTPFANQ
Н.	sapiens	MAPRKGKEKKEEQVI	SLGPQVAEGENVFGV	CHIFASENDTEVEVT	DLSGKETICRVTGGM	KVKADRDESSPYAAM
D.	melanogaster	MAPRKAKVQKEEVQV	QLGPQVRDGEIVFGV	AHIYASFNDTFVHVT	DLSGRETIARVTGGM	KVKADRDEASPYAAM
s.	pombe	MAT	NVGPQIRSGELVFGV	AHIFASENDTEVHIT	DLTGKETIVRVTGGM	KVKTDRDESSPYAAM
s.	cerevisiae	МА	NDLVQARDNSQVFGV	ARIYASENDTEVEVT	DLSGKETIARVTGGM	KVKADRDESSPYAAM
s.	solfataricus		MSSRREIRWGI	AHIYASQNNTLLTIS	DLTGAEIISRASGGM	VVKADREKSSPYAAM
м.	jannaschii		MAEQKKEKWGI	VHIYSSYNNTIIHAT	DITGAETIARVSGGR	VTRNORDEGSPYAAM
	-					
т.	thermophilus	LAALDAAKKAMAYGM	QSVDVIV <mark>R</mark> G	TGAGREQAIRALQ	ASGLQVKSIVDDTPV	PHNGCRPKKKFRKAS-
л.	aeolicus	LAAQKAMKEAKEHGV	QEVEIWVKG ·	PGAGRESAVRAVF	ASGVKVTAIR <mark>DVT</mark> PI 1	PENGCRPPARERV
Ρ.	aeruginosa	VAAERAGQAALEYGL	KNLDVNVKG	PGPGRESAVRALN	ACGYKIASITDVTPI	PENGCRPPKKRRV
Ε.	coli	VAAERCADAVKEYGI	KNLEVMVKG	PGPGRESTIRALN	AAGFRITNITDVTPI	PENGCRPPKKRRV
н.	sapiens	LAAQDVAQRCKELGI	TALHIKLRATGGNRT S	KTPGPGAQSALRALA	RSGMKIGRIE <mark>DVT</mark> PI :	PSDSTRRKGGRRGRRL
D.	melanogaster	LAAQDVAEKCKTLGI	TALHIKLRATGGNKT B	KTPGPGAQSALRALA	RSSMKIGRIEDVTPI 1	PSDSTRRKGGRRGRRL
s.	pombe	LAAQDAAAKCKEVGI	TALHIKIRATGGTAT	KTPGPGAQAALRALA	RAGMRIGRIEDVTPI	PTDSTRRKGGRRGRRL
s.	cerevisiae	LAAQDVAAKCKEVGI	TAVEVKIRATGGTRT	KTPGPGGQAAL <mark>RAL</mark> A	RSGLRIGRIEDVTPV 1	PSDSTRKKGGRRGRRL
s.	solfataricus	LAANKAASDALEKGI	MALHIKVRAPGGYGS	KTPGPGAQPAIRALA	RAGFIIGRIEDVTPI	PHDTIRRPGGRRGRRV
м.	jannaschij	OAAFKLAEVLKERGI	ENIHIKVRAPGGSGO	KNPGPGAOAAIRALA	RAGLRIGRIEDVTPV 1	PHDGTTPKKRFKK

### Multiple vs pairwise

Alignments should put together bases/amino acids that are related by evolution—roughly corresponds to being in the same structural and functional position

**Better Score** 

Correct evolutionary history

## Multiple Sequence Alignment: Approaches

- Optimal Global Alignments Dynamic programming
  - Generalization of Needleman-Wunsch
  - Find alignment that maximizes a score function
  - Computationally expensive: Time grows as product of sequence lengths
- Global Progressive Alignments Match closely-related sequences first using a guide tree

## What is an optimal multiple alignment

- Sum of pairs (SOP)
- Score of multiple alignment

 $= \sum_{i < j} \text{score}(S_i, S_j)$ 

- score(S<sub>i</sub>,S<sub>j</sub>) = score of induced pairwise alignment
- The alignment of si with sj induced by M is generated as follows
  - Remove from M all rows except i and j
  - Remove all columns that contain only blanks

## Can be solved by dynamic programming

- The two-sequence alignment algorithm can be generalized to any number of sequences.
- As for two sequences, divide possible alignments into different classes, depending on how they end.
- E.g., for three sequences X, Y, W define
   C[i,j,k] = score of optimum alignment
   among X[1..i], Y[1..j], W[1..k]

## Dynamic programming for MSA with k sequences

- K-dimensional "matrix"
- There are n<sup>k</sup> cell corners in the cube
- For each corner, we need to look at 2<sup>k</sup>-1 other corners – Together: O(2<sup>k</sup> n<sup>k</sup>) computations
- Example: 6 sequences of length 100 require 6.4X10<sup>13</sup> calculations
- Implementations (e.g., WashU MSA 2.1) use tricks and only search subset of dynamic programming table
  - Even this is expensive. E.g., Baylor CM Search launcher limits MSA to 8 sequences of 800 characters and 10 minutes processing time





## Problem with sum of pairs

- Alignment should reflect the evolutionary process
- Alignment score should be related to the number of evolutionary evens
- Sum of pairs overcounts alignments
- Too much weight for evolutionary distant pairs



CT\_TGC\_A GT\_TGACA GT\_TGTTA GTATTTCT GTATTTGA

## Progressive multiple alignment

- Align most closely related sequences first, how to decide the order
  - Ideally we would follow the evolutionary history--phyolgeny
  - Need and MSA to infer evolutionary history
  - Compute phylogeny that close enough



#### Example



A	PEEKSAV_TALWG_KVNVDEYGG
в	GEEKAAV_LALWD_KVNEEEYGG
С	PADKTNVKAA_WG_KVGAHAGEYGA
D	AADKTNVKAA_WS_KVGGHAGEYGA
E	AATNVKTA_WSSKVGGHAPAA

- C PADKTNVKAAWG\_KVGAHAGEYGA D AADKTNVKAAWS\_KVGGHAGEYGA E AA TNVKTAWSSKVGGHAPA A
- A PEEKSAVTALWGKVNVDEYGG B GEEKAAVLALWDKVNEEEYGG
- D AADKTNVKAAWSKVGGHAGEYGA
- C PADKTNVKAAWGKVGAHAGEYGA

## Clustalw

- Widely used progressive alignment approach
- Basic flow
  - Find pairwise scores
  - Build guide tree from pairwise distances (neighbor joining algorithm, discussed later)
  - Progressively align according to guide tree
- Need a way to score aligning partial alignments
  - Weighted average of pairwise scores
  - Weights correct for unequal sampling across evolutionary distance



## Gaps in Clustalw

- Opening and extension penalties depend on
  - score matrix
  - sequence similarity,
  - sequence length,
  - position of gaps
  - residues at gaps –gaps cost less in a hydrophilic region
- Gaps should cost more if they break up a structural element and less if they are in a loop
- For further details, see Thompson *et al.*, NAR 1994, 22:4673 or Methods in Enz. 1996, 266:article 22.



QLS<u>GEEK</u>AAVLALWGKVN--EEEVGGEALGRLLVVYPWTQRFFESFGDL VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLS

## Problem with progressive alignment



SeqA GARFIELD THE LAST SeqB GARFIELD THE FAST SeqC GARFIELD THE VERY SeqD THE FAST CAT FA-T CAT
---

CORREC	T (Score=	24)				
SeqA SeqB SeqC SeqD	GARFIELD GARFIELD GARFIELD	THE THE THE THE	LAST FAST VERY	FA-T  FAST FA-T	CAT CAT CAT CAT	

Solution: iterative refinement

#### BaliBase: Reference MSA based on structure



Aligner	Performance*	Time		
DIALIGN	57.2	12 h, 25 min		
CLUSTALW	58.9	2 h, 57 min		
T-Coffee	63.6	144 h, 51 min		
MUSCLE	64.8	3 h, 11 min		
MAFFT	64.8	2h,36min		
ProbCons	66.9	19 h, 41 min		
ProbCons-ext	68.0	37 h, 46 min		

Do et al, Genome Research, 2005

### Which program to choose



Do and Katoh, 2008

## Multiple Alignments Summary

- Even below the 10-20% identity twilight zone, the best programs correctly align 47% of residues on average
- Iterative algorithms are superior, but with a large trade-off in use of computational resources
- Global generally performs better than local
- No single 'best' program exists
- For reviews, see:
  - P. Briffeuil *et al.*, **Bioinformatics** 1998, **14**:357
  - J. D. Thompson *et al.*, **NAR** 1999, **27**:2682

#### Complex dependence structure



Slide credit: Dannie Durand