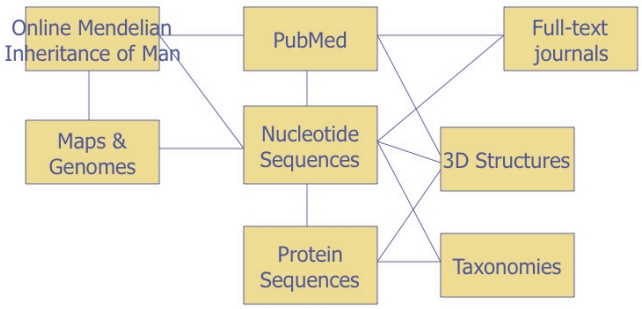
1.2

Databases and heuristic algorithms

Databases at the NCBI



National Center for Biotechnology

Sequence databases

(1) Store any sequences and fragments that have been found
(2) Unique accession key (arbitrary number)
(3) Attributes such as source, species, etc.
(4) One attribute is the sequence string

Example GenBank entry (1)

```
XX
AC   X04751;
XX
SV   X04751.1
XX
DT   07-JUN-1987 (Rel. 12, Created)
DT   10-FEB-1999 (Rel. 58, Last updated, Version 5)
XX
DE   Rabbit alpha-1-globin gene to theta-1-globin pseudogene region
XX
KW   alpha-1-globin; alpha-globin; globin; pseudogene; repetitive sequence;
KW   tandem repeat; theta-1-globin; theta-globin.
XX
OS   Oryctolagus cuniculus (rabbit)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Lagomorpha; Leporidae; Oryctolagus.
XX
RN   [1]
RP   1-4028
RA   Hardison R.C.;
RT   ;
RL   Submitted (02-FEB-1987) to the EMBL/GenBank/DDBJ databases.
RL   Hardison R.C., Pennsylvania State University, Althouse Laboratory,
RL   University Park, Pennsylvania 16802, USA.
```


Example GenBank entry (2)

```
XX
RN   [2]
RP   1-4028
RX   MEDLINE; 86085923.
RA   Cheng J.-F.F., Raid L., Hardison R.C.;
RT   "Isolation and nucleotide sequence of the rabbit globin gene cluster
RT   psi-zeta-alpha-1-psi-alpha: Absence of a pair of alpha-globin genes
RT   evolving in concert";
RL   J. Biol. Chem. 261:839-848(1986).
XX
DR   EPD; EP11096; OC_HBA.
DR   SWISS-PROT; P01948; HBA_RABIT.
XX
CC   Submitted data [2] include some corrections to published seq. [1].
CC   Referring to the authors the sequence from pos. 50 to 70 may not
CC   be completely accurate due to reading problems of the sequencing
CC   gels.
CC   Theta-1 pseudogene was formerly called psi alpha.
CC   Data kindly reviewed (15-Jun-1987) by Hardison R.C.
XX
```


Example GenBank entry (3)

```
FH   Key             Location/Qualifiers
FH
FT   source          1..4028
FT                   /db_xref="taxon:9986"
FT                   /organism="Oryctolagus cuniculus"
FT   precursor_RNA   150..861
FT                   /note="primary transcript od alpha-1-globin"
FT   exon            150..280
FT                   /number=1
FT   CDS             join(186..280,358..562,646..774)
FT                   /db_xref="SWISS-PROT:P01948"
FT                   /product="alpha-1-globin"
FT                   /protein_id="CAA28447.1"
FT                   /translation="MVLSPADKTNIKTAWEKIGSHGGEYGAEAVERMFLGFPTTKTYFP
FT                   HFDFTHGSEQIKAHGKKVSEALTKAVGHLDDLPGALSTLSDLHAHKLRVDPVNFKLLSH
FT                   CLLVTLANHHPSEFTPAVHASLDKFLANVSTVLTSKYR"
```

Example GenBank entry (4)

```
FT   intron          281..357
FT                   /number=1
FT   exon            358..562
FT                   /number=2
FT   intron          563..645
FT                   /number=2
FT   exon            646..861
FT                   /number=3
FT   polyA_signal    841..846
FT   polyA_site      861..861
FT   repeat_region   1542..1675
FT                   /note="region of 5 x 25bp tandem repeat 1"
FT   repeat_region   3067..3133
FT                   /note="region of 7 tandem repeat 2 (9-10bp)"
FT   CDS             3139..3744
FT                   /pseudo
FT                   /product="theta-1-globin"
FT   polyA_signal    3803..3808
FT   polyA_site      3818..3818
FT                   /note="put. polyA site (found by homology to alpha-1)"
XX
```


Example GenBank entry (5)

```
SQ   Sequence 4028 BP; 685 A; 1359 C; 1310 G; 674 T; 0 other;
     gcggggccgg gtcccaggca gacgccgcga gggcgccccc agcggtggcg gccgccgccg        60
     cgccccgccg cgccggccaa tgagcggggc cccgctgggc gtgcccgcag cacctcgggc       120
     ttaaaagcgc cgcgcagtct gggctccgca cacttctggt ccagtccgac tgagaaggaa       180
     ccaccatggt gctgtctccc gctgacaaga ccaacatcaa gactgcctgg gaaaagatcg       240
     gcagccacgg tggcgagtat ggcgccgagg ccgtggagag gtgaggaccc ccgccccgcc       300
     ccgccccgcc cgagcccgcc ggcgccgcgc cccgctcacg gcctcctgtc cccgcaggat       360
     gttcttgggc ttccccacca ccaagaccta cttcccccac ttcgacttca cccacggctc       420
     tgagcagatc aaagcccacg gcaagaaggt gtccgaagcc ctgaccaagg ccgtgggcca       480
     cctggacgac ctgcccggcg ccctgtctac tctcagcgac ctgcacgcgc acaagctgcg       540
     ggtggacccg gtgaatttca aggtgagccc gcagccccggc tgggagcgtc gcgggggtcg      600
     gcggtccccg accacaccca ccgacgtccg cccctctctc tgcagctcct gtcccactgc       660
     ctgctggtga ccctggccaa ccaccacccc agtgaattca cccctgcggt gcacgcctcc       720
     ctggacaagt tcctggccaa cgtgagcacc gtgctgacct ccaaatatcg ttaagctgga       780
     gcctgggagc cggcctggcc ctccgccccc cccacccccg cagcccaccc ctggtctttg       840
     aataaagtct gagtgagtgg ccgacagtgc ccgtggagtt ctcgtgacct gaggtgcagg       900
     gccggcctag ggacacgtcc gtgcacgtgc cgaggccccc tgtgcagctg caagggacag       960
     gagtgggcaa ccggctggtt ccttccttcc tgcttgcaag tccacgaggg gctgctgaaa      1020
     gaacccccca cacacacatg cacacactcg tgccactcgg ctgcctccag cctgggtccc      1080
     ....
```

Finding related sequences

(1) Find related sequences to a target sequence
- Process all entries of the database and do matching
- Computationally expensive

(2) Faster alignment algorithms needed
- Optimality cannot be guaranteed
- Focus on matching ungapped segments
- FAST and BLAST

BLAST

(1) Segment pairs
- Segment pair: Two aligned subsequences without gaps
- Find all high-scoring segment pairs between two sequences
- Similar to a gapped sequence without scoring gaps
- Heuristics focus on locally conserved/related sequences

(2) Steps of the algorithm
- Find all words (e.g. length = 4 characters) that match somewhere in the query sequence with a score > T
- Find occurrences of these words (seeds) in the comparison string
- Extend seeds in both directions until score drops below a fraction of the maximum so far
- Report all segment pairs with score > S

Further parameter considerations

(1) Low-complexity regions
- Regions of low variation (only few different amino acids with high repetition rate)
- Can produce high scoring hits
- Biologically „assumed" to be irrelevant/non-functional
- Can be filtered (substituted by X) before query

(2) Different substitution matrices
- Probabilities calculated for specific number of evolutionary steps
- PAM60 means 60 changes of the sequence (vs. PAM120)
- Higher numbers for dissimilar sequences, lower numbers for related sequences

FAST

(1) Find high scoring offset
- E.g., s = H A R F Y A A Q I V L

(2) Lookup tuples (ktup = 1 or 2) and record offset
- A (2, 6, 7), F (4), H (1), I (9), L (11), Q (8), R (3), V (10), Y (5)

(3) Scan database string and count offsets
- E.g., t = V D M A A Q I A
- Pos 1 (V): 10 – 1 = 9: [9]
- Pos 4 (A): 2 – 4 = -2, 6 – 4 = 2, 7 – 4 = 3: [-2, 2, 3]
- Pos 5 (A): [-3, 1, 2], Pos 6 (Q): [2], Pos 7 (I): [2], Pos 8 (A): [-6, -2, -1]
- Offset 2 occurs 4 times

(4) Called „diagonal method"
- Find diagonal in the dynamic programming matrix

FAST (2)

(1) Heuristically join tuples into regions (gapless alignments)
- E.g., A A Q I with offset 2

(2) Rescore regions with substitution matrix
- Best = initial score

(3) Finally recalculate using dynamic programming restricted to a band around the diagonal found