

Bio 595

Computers in Biomedical Research Fall 2003 Slides for Wednesday, December 12

PDB file format

PDB files consist of lines of 80 columns; lines begin with a predefined record name and end with a newline.

Different record types have different fields defined:

Primary structure group:

DBREF a reference to the entry in the sequence database

SEQADV conflicts between the PDB and the named sequence database

SEQRES primary sequence of backbone residues

MODRES modification to standard residues

Databases referenced in PDB files

BMRB	BioMagResBank
BLOCKS	BLOCKS
EMBL	European Molecular Biology Laboratory
GB	GenBank
GDB	Genome Data Base
NDB	Nucleic Acid Database
PROSIT	PROSITE
PDB	Protein Data Bank
PIR	Protein Identification Resource
SWS	SWISS-PROT
TREMBL	TREMBL

PDB Files

PDB makes it easy to collect record types: they all start with the same keyword at the beginning of each line

Note subroutine extractSEQRES.

Subroutine iub3to1 converts 3 character codes into 1 character codes.

Other data provided in PDB: detailed structural information, including atomic coordinate data (from x-ray crystallographic data, NMR data); these allow prediction of secondary structure data

BLAST

Basic Local Alignment Search Tool –tests a query sequence against a library of known sequences to find a similarity. It's a string-matching program. Similarities in strings indicates homology (meaning sequences are related evolutionarily), and can be expressed by the percent identity. Can also be measured by the 'degree of conservation', looking for redundant (equivalent) codons or between amino acid residues with similar properties that don't alter protein function.

BLAST programs do query-to-database comparisons for nucleotide-nucleotide, protein-nucleotide, protein-protein, nucleotide-protein.

Getting BLAST: <http://www.ncbi.nlm.nih.gov/BLAST/>

Site allows download of BLAST, plus tutorials on its use, plus comprehensive set of databases

BLAST statistics

BLAST output reports scores and statistics based on a raw score S. Here S is a measure of similarity and the size of the match.

BLAST output lists hits ranked by their E (expect) value, which measures the chances that the string matching (allowing for the gaps) could occur in a randomly generated database of the same size and composition. The closer E is to zero, the less likely the sequence occurred by chance (i.e., the better the match).

In approximate terms, values less than 1 are a solid hit; values less than about 10 are probably worth examining further.

BLAST output file format

BLASTN 2.1.3 [Apr-11-2001]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Zhigui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1990). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.
 RID: 99153553-27495-9092
 Query: (400 letters)

Database: nt
 868,831 sequences; 3,298,558,333 total letters

Sequences producing significant alignments:	Score	E
	(bits)	Value
dbj AB031069.1 AB031069 Homo sapiens PCCK1 mRNA for protein cont...	793	0.0
ref MP_014393.1 Homo sapiens Cpc binding protein (CCBP), mRNA	779	0.0
gb AF149758.1 AF149758 Homo sapiens Cpc binding protein (CCBP) m...	779	0.0
ref JX_008699.1 Homo sapiens Cpc binding protein (CCBP), mRNA	765	0.0
emb AL138862.1 HSMB01830 Homo sapiens mRNA; cDNA DKF2p434f174 (f...	450	e-124
emb A312339.1 HS412339 Homo sapiens Cpc island sequence, subcl...	446	e-123
emb A312340.1 HS412340 Homo sapiens chromosome 18 Cpc island D...	405	e-123
dbj AK010337.1 AK010337 Mus musculus ES cells cDNA, RIKEN full-l...	234	3e-59
dbj AK017941.1 AK017941 Mus musculus adult male thymus cDNA, RIK...	210	5e-52
gb AC009750.2 AC009750 Drosophila melanogaster, chromosome 2L, f...	86	0.017
gb AE003580.2 AE003580 Drosophila melanogaster genomic scaffold...	66	0.027
ref NC_003905.1 Leishmania major chromosome 1, complete sequence	40	1.0
gb AE001274.1 AE001274 Leishmania major chromosome 1, complete s...	40	1.0
gb AC008259.2 AC008259 Drosophila melanogaster, chromosome 3R, f...	38	4.1
gb AC018662.3 AC018662 Human Chromosome 7 clone RP11-339C9, comp...	38	4.1

A perfect hit in BLAST

ALIGNMENTS

>dbj|AB031069.1|AB031069 Homo sapiens PCCK1 mRNA for protein containing CXK domain 1,

complete cds
 Length = 2487

Score = 793 bits (400), Expect = 0.0

```

Query: 1  agatggcggcgtgaggggtcttgggggcttagggccacctactggtttgcagcg 60
          |||
Sbjct: 1  agatggcggcgtgaggggtcttgggggcttagggccacctactggtttgcagcg 60

Query: 61  agacgacgatggggcctgcgaataggagtagctgctgggagcgtgactagaagc 120
          |||
Sbjct: 61  agacgacgatggggcctgcgaataggagtagctgctgggagcgtgactagaagc 120

Query: 121  gaagtgtgtggggcctttgcaaccgctgggaccccgagtggtctgtcaggtt 180
          |||
Sbjct: 121  gaagtgtgtggggcctttgcaaccgctgggaccccgagtggtctgtcaggtt 180

Query: 181  cgcgggtcgttggcggggctgtgaggagtagcgcgggagcggagataggaggat 240
          |||
Sbjct: 181  cgcgggtcgttggcggggctgtgaggagtagcgcgggagcggagataggaggat 240
  
```

A good hit in BLAST

>dbj|AK017941.1|AK017941 Mus musculus adult male thymus cDNA, RIKEN full-length enriched library, clone:5830420C16, full insert sequence
 Length = 1461

Score = 210 bits (106), Expect = 5e-52
 Identities = 151/166 (90%)
 Strand = Plus / Plus

```

Query: 235  ggagatggttcagaccagagcctccagatccggggaggacagcaagtcgagaatggg 294
          |||
Sbjct: 1048  ggagatggttcagaccagagcctccagatccggggaggacagcaagtcgagaatggg 1107

Query: 295  gagaagtcgccatctactgcatctgcccgaaccgacatcaattgcttcatgatcgg 354
          |||
Sbjct: 1108  gagaagtcgccatctactgcatctgcccgaaccgacatcaattgcttcatgatgga 1167

Query: 355  tgtgacaactgcaatgagtggttccatgggagctcatccgatca 400
          |||
Sbjct: 1168  tgtgacaactgcaatgagtggttccatgggagctcatccgatca 1213
  
```