2.3

Protein databases

Protein databases

(1) Sequence and structure
- E.g., SwissProt
- Sequence information as in other sequence databases
- Structure information as PDB (protein data base) file
- Patterns and families

(2) Protein interactions
- E.g., BIND (biomolecular interactions data base)
- Generalized functional information

(3) Accession keys
- Unique identification of protein/database entry
- Should neither change nor carry information
- For reference between entries and databases

SwissProt entry

NiceProt View of SWISS-PROT: P32905

Printer-friendly view | Quick BlastP search

[General] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

General information about the entry	
Entry name	RS0A_YEAST
Primary accession number	P32905
Secondary accession numbers	None
Entered in SWISS-PROT in	Release 27, October 1993
Sequence was last modified in	Release 30, October 1994
Annotations were last modified in	Release 37, December 1998

SwissProt entry (2)

Name and origin of the protein	
Protein name	40S ribosomal protein S0-A
Synonym	Nucleic acid-binding protein NAB1A
Gene name	RPS0A or NAB1A or NAB1 or YST1 or YGR214W
From	Saccharomyces cerevisiae (Baker's yeast) [TaxID: 4932]
Taxonomy	Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.

References

[1] SEQUENCE FROM NUCLEIC ACID.
Miles J., Formosa T.G.;
Submitted (MAR-1992) to the EMBL/GenBank/DDBJ databases.

[2] SEQUENCE FROM NUCLEIC ACID.
STRAIN=S288c;
MEDLINE=97435481; PubMed=9290212; [NCBI, ExPASy, EBI, Israel, Japan]
Rieger M., Brueckner M., Schaefer M., Mueller-Auer S.;
"Sequence analysis of 203 kilobases from Saccharomyces cerevisiae chromosome VII.";
Yeast 13:1077-1090(1997).

2-37

Comments

- *FUNCTION*: BINDS DNA. REQUIRED FOR THE ASSEMBLY AND/OR STABILITY OF THE 40S RIBOSOMAL SUBUNIT.
- *MISCELLANEOUS*: THERE ARE TWO GENES FOR S0 IN YEAST.
- *SIMILARITY*: BELONGS TO THE S2P FAMILY OF RIBOSOMAL PROTEINS.

Copyright

Cross-references

EMBL	M88277; AAB05643.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] Z72999; CAA97241.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	S42143; S42143.
SWISS-2DPAGE	P32905; YEAST.
SGD	S0003446; RPS0A.
GeneCensus	P32905; YGR214W.
InterPro	IPR001865; Ribosomal_S2. Graphical view of domain structure.
Pfam	PF00318; Ribosomal_S2; 1.
PRINTS	PR00395; RIBOSOMALS2.
TIGRFAMs	TIGR01012; Sa_S2_E_A; 1.
PROSITE	PS00962; RIBOSOMAL_S2_1; 1. PS00963; RIBOSOMAL_S2_2; 1.
ProDom	[Domain structure / List of seq. sharing at least 1 domain].
BLOCKS	P32905.
ProtoMap	P32905.
PRESAGE	P32905.
DIP	P32905.
ModBase	P32905.

2-38

SwissProt entry (4)

Keywords

Ribosomal protein; DNA-binding; Acetylation; Multigene family.

Features

Key	From	To	Length	Description
INIT_MET	0	0		
MOD_RES	1	1		ACETYLATION.

 Feature table viewer

Sequence information

Length: 251 AA | Molecular weight: 27893 Da | CRC64: 4FF263575B82C75A [This is a checksum on the sequence]

```
        10         20         30         40         50         60
         |          |          |          |          |          |
SLPATFDLTP EDAQLLLAAN THLGARNVQV HQEPYVFNAR PDGVHVINVG KTWEKLVLAA

        70         80         90        100        110        120
         |          |          |          |          |          |
RIIAAIPNPE DVVAISSRTF GQRAVLKFAA HTGATPIAGR FTPGSFTNYI TRSFKEPRLV

       130        140        150        160        170        180
         |          |          |          |          |          |
IVTDPRSDAQ AIKEASYVNI PVIALTDLDS PSEFVDVAIP CNNRGKHSIG LIWYLLAREV

       190        200        210        220        230        240
         |          |          |          |          |          |
LRLRGALVDR TQPWSIMPDL YFYRDPEEVE QQVAEEATTE EAGEEEAKEE VTEEQAEATE

       250
         |
WAEENADNVE W
```

P32905 in FASTA format

2-39

BIND entry

Interaction

Interaction ID: 6155

Accession date: Sep 5, 2001
Description: Tyrosine phosphorylated Gab1 recruits PI3K by direct interaction with the p85 subunit

Molecule A

Gab1
Description: Grb2-Associated Binder-1. A docking protein that contains a PH domain, several proline-rich stretches and multiple tyrosine phosphorylation sites which are SH2
Molecule Type: Protein
GI: 4503851 (NCBI) (SEQHOUND) (BIND)
Molecule origin: Organismal
Organism: Homo sapiens

BIND entry (2)

Molecule B

PI3K p85-alpha
Description: Phosphatidylinositol 3-kinase, p85 subunit {alpha}.
Molecule Type: Protein
GI: 105122 (NCBI) (SEQHOUND) (BIND)
Molecule origin: Organismal
Organism: Homo sapiens

Visualize Interaction!

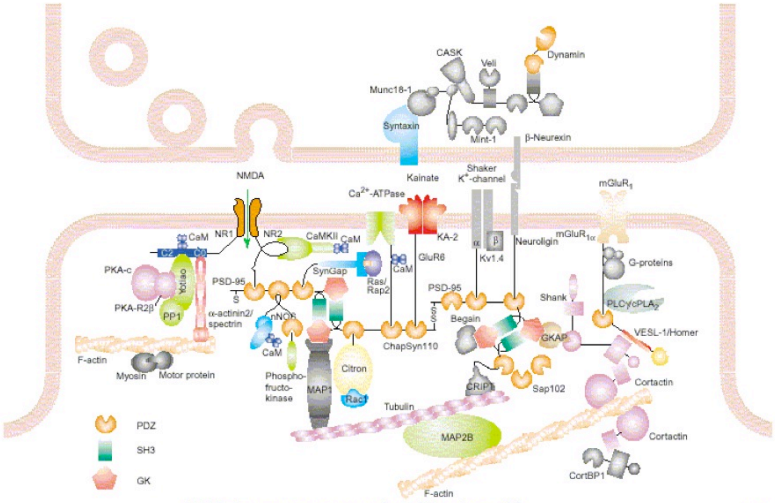
View other information?

Main Info	Publications	ASN.1 XML
Cellular Place	Experimental Condition	Conserved Sequence
N/A	N/A	N/A
Binding Sites	Chemical action	Chemical State
N/A	N/A	N/A

Comments and suggestions to: < info@bind.ca >

BIND Homepage

Protein complexes



NMDA receptor complex (Husi et al.)

2.4

Protein structure prediction

Protein structure prediction

(1) Homology-based
- Fragment approaches: finding aligned segments without gaps (SCR structurally conserved regions, e.g., to identify helices or sheets from known sequences) and loops (SVR structurally variable regions) to match other known segments with least variation

(2) Threading
- Generation and checking of many different (rough) alignments
- Scoring, e.g. with probability distribution of distances between amino acid pairs (position and sequence number)

(3) Conformational energy
- Generate conformations
- Minimize energy function
- Not efficient!/Local minima!

Metropolis algorithm

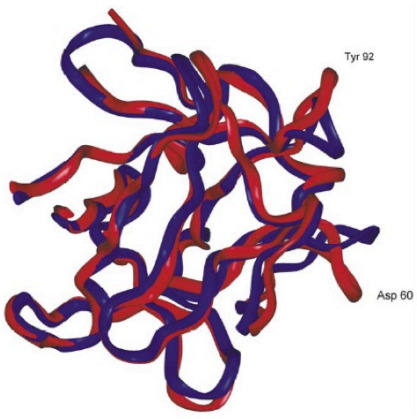
(1) Start
- Any molecule conformation x with energy E (x)

(2) Loop
- Randomly disturb x into x‘
- If E (x‘) < E (x) continue loop with x‘
- Else randomly choose whether to continue with x or x‘ (probability depending on a virtual temperature T)

(3) Can escape local minima
- If temperature is high, randomly move away from local minima with some finite probability

(4) Simulated annealing
- Start with a high temperature
- Slowly lower the temperature to move into minimum
- Guaranteed success with slow temperature reduction
- But usually faster reduction required

Example structure prediction



Red is prediction by SwissModel (RMS = 1,54Å)

2.5

Protein identification by mass spectrometry

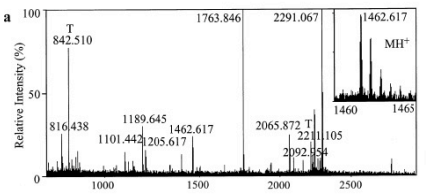
Mass spectrometry

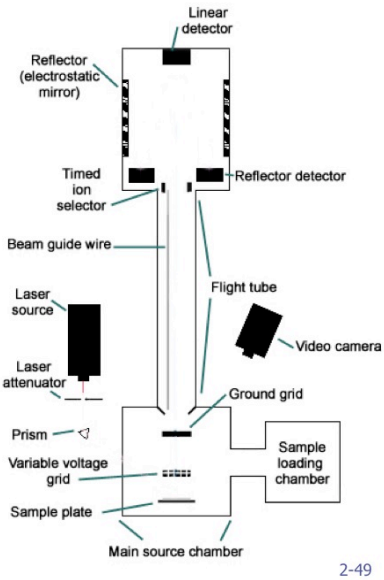
(1) Mass spectrometry separates according to mass

- E.g., TOF (time of flight): peptides are ionized and accelerated in an electric field
- Detector produces a peak for several flight times, flight time correlates with mass

Mass spectrometer

TOF Time of Flight

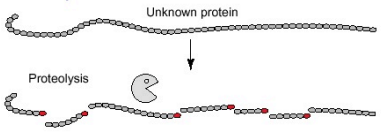


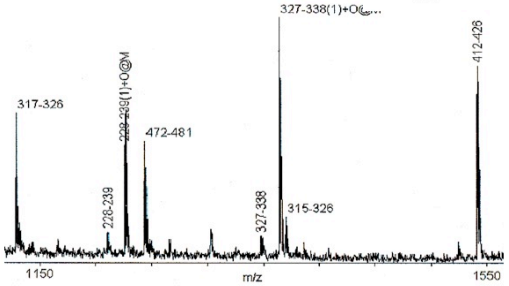


Peptide mass fingerprinting

(1) Proteins are split into peptides by sequence-specific proteases

- Always break at the same amino acid or pattern
- For a given sequence, a unique set of peptides results
- Mass of all peptides is measured





Peptide mass fingerprinting

(1) Database search
- Mass spectrum is compared with theoretical spectra for proteins in a database
- Best match is obtained

(2) Match
- Peak from spectrum equals one theoretical peptide
- Within mass tolerance/accuracy
- Multiple matches from the spectrum possible (counts only as one)

(3) Shared peak count
- Highest number of matches

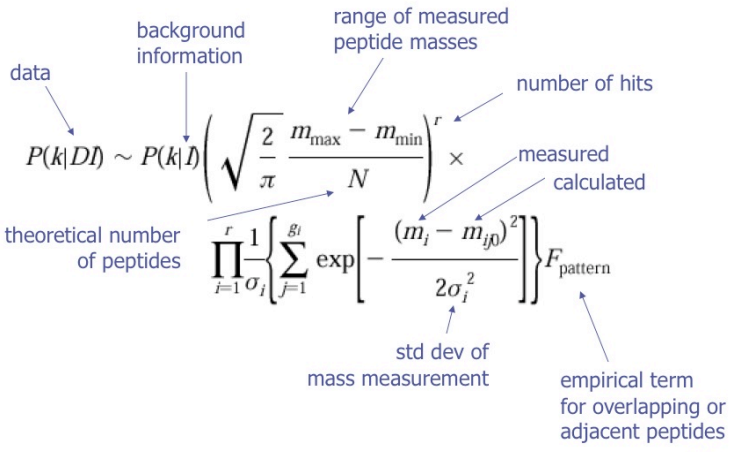
Problems with shared peak count

(1) Not all peptides occur in the mass spectrum
- low abundance, further fragmentation, incomplete ionisation, phosphorylation, etc.

(2) Is the protein really in the database?
- spurious matches with proteins from other organisms, mutations

(3) Long proteins tend to be preferred
- as the are more opportunities for matches

Statistical approach: ProFound



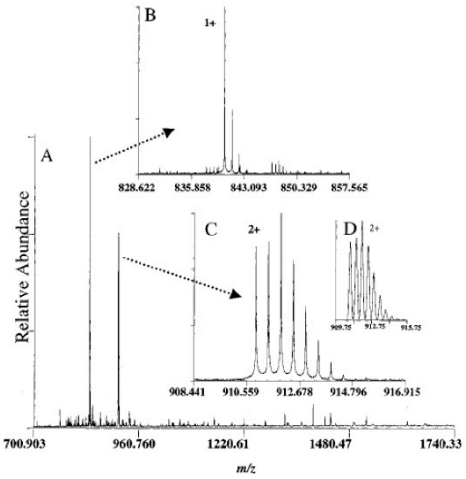
$$P(k|DI) \sim P(k|I)\left(\sqrt{\frac{2}{\pi}}\,\frac{m_{\max}-m_{\min}}{N}\right)^{r} \times \prod_{i=1}^{r}\frac{1}{\sigma_i}\left\{\sum_{j=1}^{g_i}\exp\left[-\frac{(m_i-m_{ij0})^2}{2\sigma_i^2}\right]\right\}F_{\text{pattern}}$$

Single peptide identification

(1) Experimental constraints should make database search result unique

- High resolution (high accuracy) MS restricts hits
- Cysteine-containing peptides only (rare):
  Cysteine tag removes all other peptides
- Known organism (limited number of genes/proteins)

High resolution mass spectrometry



Example: Isobaric peptides in yeast

protein name[a]	protein mass[b]	peptide isobars	peptide mass	error (ppm)
laminin fragment	2 426.271	***R***YVVLP***R***PVCFE***K***	1604.886[c]	0.189
TPI1	26 778.962	FLAS***K***LGD***K***AASEL***R***	1604.889	1.747
YPR143W	52 661.268	D***KKR***T***RK***NAEFG***R***	1604.886	0.07
YDR428C	52 785.179	NLYDAVSNIT***R***LV***K***	1604.889	1.747
TAP42	31 135.369	IELFQ***K***N***K***EIST***K***	1604.889	1.747
CYC2	37 692.826	VQL***K***IFETD***R***QT***K***	1604.889	1.747
MFS1	46 151.311	ESHPVGIL***R***DLIE***K***	1604.889	1.747
YMR291W	64 850.898	EFDLL***R***SISE***K***L***R***	1604.889	1.747
YKR078W	51 708.695	I***R***TAEDEY***R***IVL***K***	1604.889	1.747
TFC6	75 311.526	D***K***IE***R***IYGLN***K***E***K***	1604.889	1.747
SSE1	77 318.483	YLA***K***EEE***KK***QAI***R***	1604.889	1.747
YBR102C	85 484.685	LDEFI***KK***NSD***K***L***R***	1604.889	1.747
STB6	88 779.841	***K***ISADLN***K***IDGLYR	1604.889	1.747
FZ01	97 746.957	E***K***NGFNIE***KK***ALS***K***	1604.889	1.747
SEC10	100 279.455	NES***K***IV***KR***VFEE***K***	1604.889	1.747
YLL005C	102 103.872	I***K***ELLFELFYY***K***	1604.885	0.234
S51441	105 161.643	HTVTEL***K***SEIHAL***K***	1604.889	1.747
PEX1	117 202.758	EEV***K***DIIE***R***HLP***K***	1604.889	1.747
RRP5	193 015.955	A***K***D***KKK***VEDLFE***R***	1604.889	1.747
DOP1	194 565.002	LTSSLSPALPAGVHQ***K***	1604.889	1.747

Only one containing cysteine

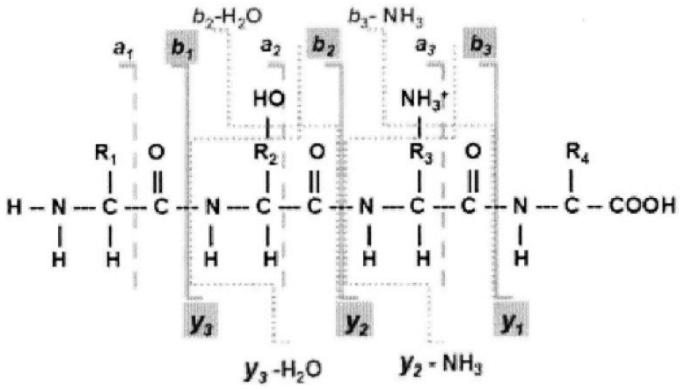
Distinction also possible through protein mass

MS/MS

(1) Subject peaks to a further MS step
- Breaking each peptide (mechanically) into ion fragments
- Measuring fragment spectrum
- Identify peptide sequence through comparison with theoretical fragmentation

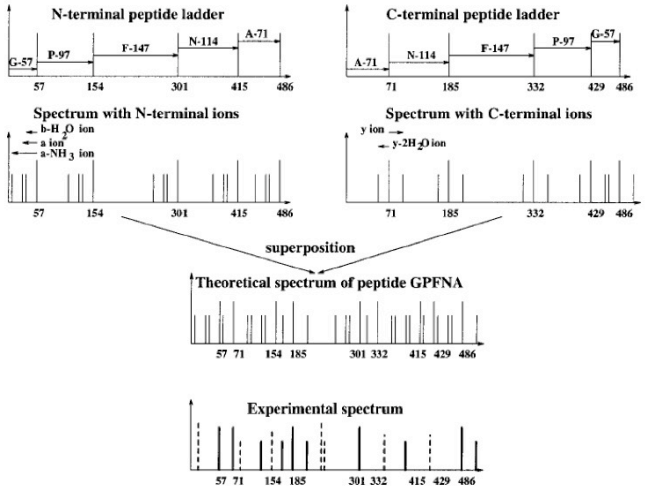
(2) For unknown organisms
- Can search among data from all organisms and even partial sequences

Typical ions



Mostly C-terminal or N-terminal ions (few middle fragments)

Experimental vs theoretical spectrum



Peptide identification problem

(1) Peptide P
- Sequence of amino acids $p_1...p_n$
- Mass m (p) = sum (m (p_i))

(2) Ion types
- $\{d_1,..,d_k\}$ numbers (weight difference)
- d-ion of partial peptide P has mass m (P) - d

(3) Spectrum
- $S = \{s_1,..,s_n\}$ is a set of (experimental) masses

(4) Problem:
- Given S, ion types and mass m
- Find a peptide of mass m with maximal match to S
- Scoring function can be shared peak count or more probabilistic

The effect of mutations

(1) A single mutation halves the shared peak count