

Building a sequence search protocol

Source textbook

“ Introduction to Bioinformatics ”

Chapter 9

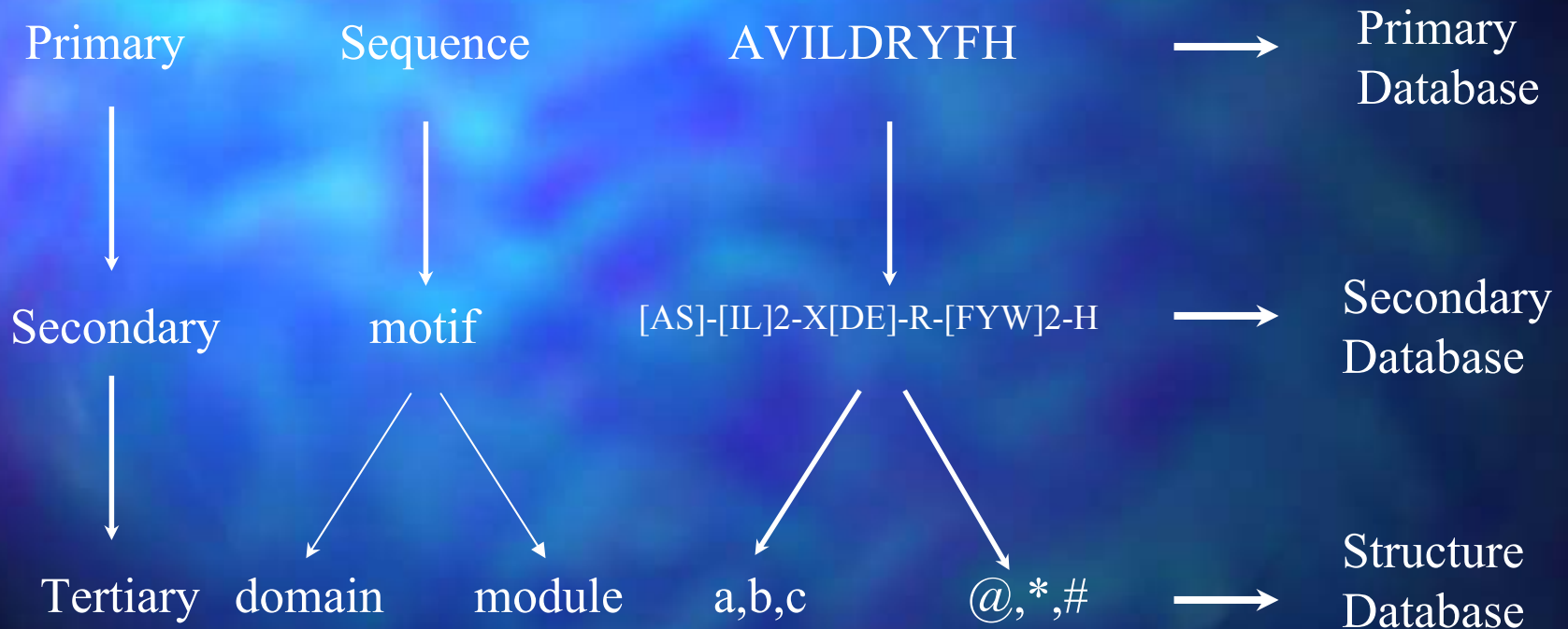
A practical approach

- Given a newly determined sequence →
 - What is my protein ?
 - To what family does it belong ?
 - What is its function ?
 - How can we explain its function in structural terms ?

A practical approach

- By searching secondary database →
 - Can get abstractions of functional
 - Structural sites characteristic of particular proteins
- Recognize patterns →
 - Allow us to infer relationships with previously characterized families
- Searching fold libraries →
 - Can get templates of known structure → possible to recognize a previously characterized fold

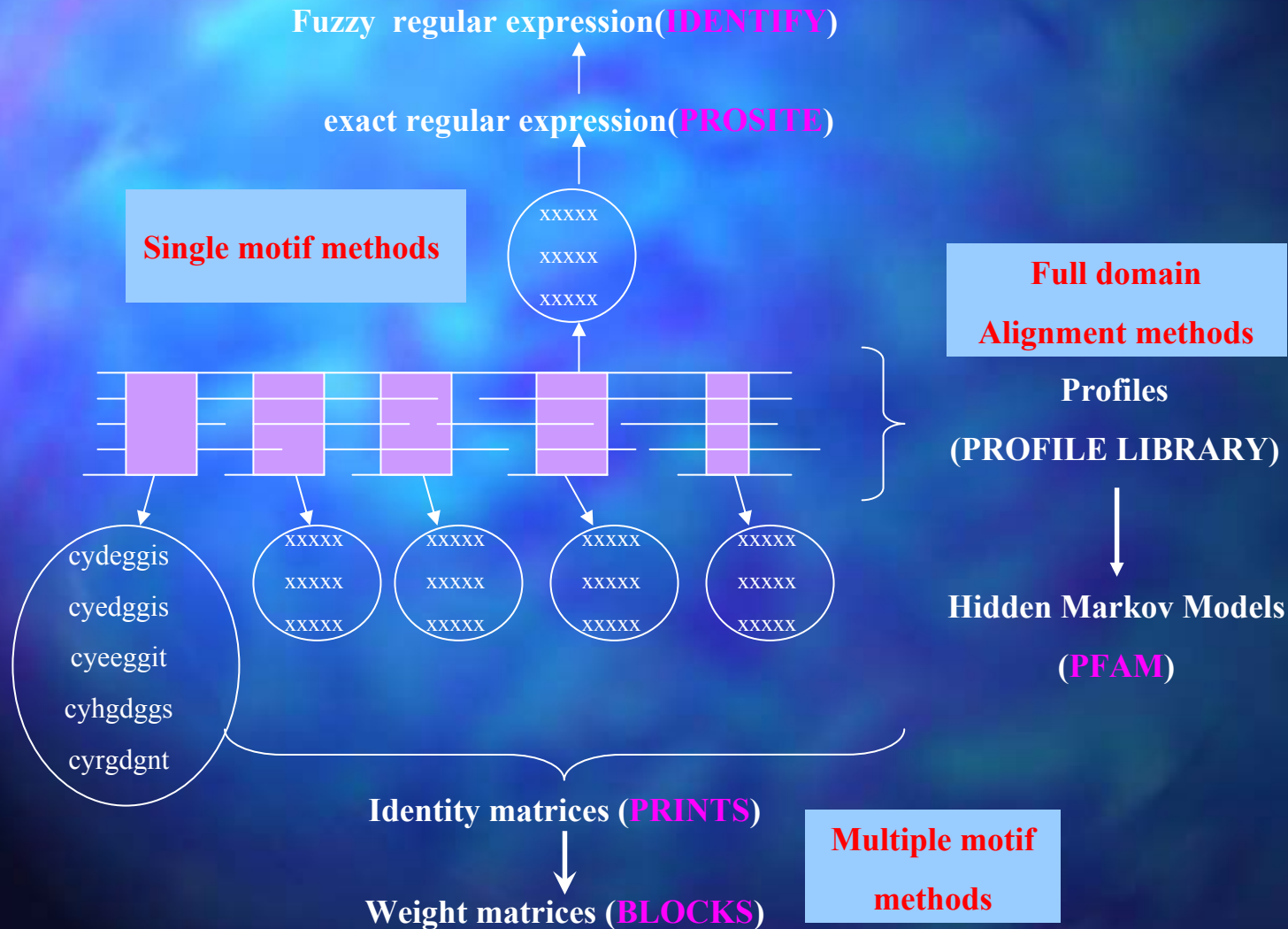
Levels of protein sequence and structural organisation



Major secondary “pattern” database

Secondary database	Primary source	Stored information
PROSITE	SWISS-PROT	Regular expressions(patterns)
Profiles	SWISS-PROT	Weighted matrices(profiles)
PRINTS	OWL*	Aligned motifs(fingerprints)
Pfam	SWISS-PROT	Hidden Markov Models(HMMs)
BLOCKS	PROSITE/PRINTS	Aligned motifs(blocks)
IDENTIFY	BLOCKS/PRINTS	Fuzzy regular expressions(patterns)

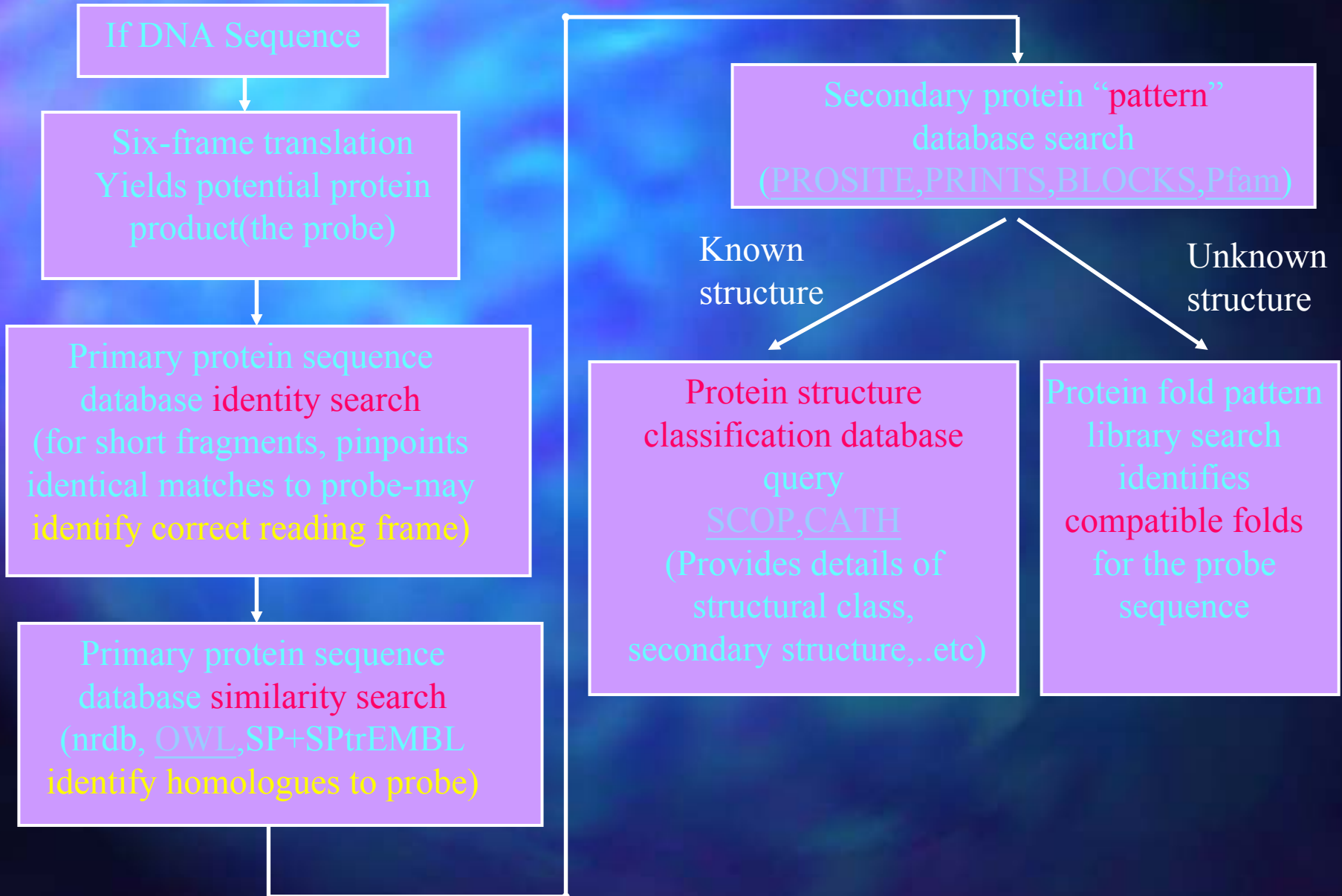
Three principal methods for building pattern databases



A practical approach

- If we cannot predict function or structure directly from sequence, but can identify homologues, and recognize sequence and fold patterns that have already been seen, how do we use this information to build a sensible search protocol for novel sequence ?

A practical approach



Searching the primary database

- Given an unknown fragment of DNA
 - ccg tac tac aac tac gct ggt gca ttc aag
- It can be translated by six-frame translator.

PYYNYAGAFK

LECTSVVVR

RTTTTLVHS

LNAPA ! L ! Y

VLQLRWCIQ

! MHQESCST

3 Forward

3 Reverse

Searching the primary database

SEQUENCE TRANSLATION - Microsoft Internet Explorer - [離線工作]

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

← 上一頁 → 搜尋 我的最愛 記錄

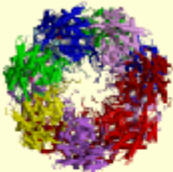
網址(D) http://www.bioinf.man.ac.uk/dbbrowser/bioactivity/nucleicfrm.html

[Contents](#)
[Commentary](#)
[Logo](#)
[Home](#)

Sequence translation & identification
From now on, it is necessary to work with several windows. Retrieve your DNA sequence using the [Translator](#). To discover the correct translation, click on **...go?** in the fragment window you find a hit.

Nucleic Acid Translator

Input your nucleic acid sequence here:



Bioinformatics
Web Practical

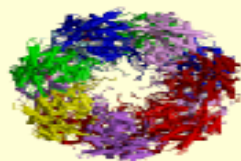
ccgtactacaactacgctggtgcattcaag

Send Query Reset

[Contents](#)
[Commentary](#)
[Logo](#)
[Home](#)

Sequence translation & identification

From now on, it is necessary to work with several windows. Retrieve your sequence in the [Sequence](#) window. To discover the correct translation, click on **...go?** in the window you find a hit.



Bioinformatics
Web Practical



Six-frame translation:

● Forward 0

10
0 PYINYAGAFK

● Forward 1

10
0 RTTTLVHS

● Forward 2

10
0 VLQLRWCIQ

● Reverse 0

10
0 LECTSVVVR

● Reverse 1

10
0 LNAPA!L!Y

● Reverse 2

10
0 !MHQRSCST

Searching the primary database

← 上一頁 → 搜尋 我的最愛 記錄

網址 http://bioinf.man.ac.uk/cgi-bin/dbbrowser/OWL/DoOWL-remote.pl

WORKLIST ENTRIES (2):

[TRFE_XENLA](#) TRANSFERRIN PRECURSOR. - XENOPUS LAEVIS (AFRICAN CLAWED FROG).
Species: XENOPUS LAEVIS (AFRICAN CLAWED FROG).
Accession: P20233
MOSKAITIS J.E., PASTORI R.L., and SCHOENBERG D.R. NUCLEIC ACIDS
RES. 18:6135-6135(1990).
GIBSON T.J. UNPUBLISHED OBSERVATIONS (MAR-1995).
01-FEB-1991 (REL. 17, CREATED)
01-NOV-1995 (REL. 32, LAST SEQUENCE UPDATE)
01-NOV-1995 (REL. 32, LAST ANNOTATION UPDATE)
-!- FUNCTION: TRANSFERRINS ARE IRON BINDING TRANSPORT PROTEINS
WHICH
CAN BIND TWO ATOMS OF FERRIC IRON IN ASSOCIATION WITH THE BINDING
OF AN ANION, USUALLY BICARBONATE. IT IS RESPONSIBLE FOR THE
TRANSPORT OF IRON FROM SITES OF ABSORPTION AND HEME DEGRADATION
TO THOSE OF STORAGE AND UTILIZATION. SERUM TRANSFERRIN MAY ALSO
HAVE A FURTHER ROLE IN STIMULATING CELL PROLIFERATION.
-!- SUBUNIT: MONOMER.
-!- DOMAIN: COMPOSED OF TWO HOMOLOGOUS DOMAINS.
-!- SIMILARITY: BELONGS TO THE TRANSFERRIN FAMILY.
-!- CAUTION: REF.1 SEQUENCE DIFFERS FROM THAT SHOWN IN POSITIONS
66
TO 79, 298 TO 332, AND 691 TO 701 DUE TO FRAMESHIFTS.
EMBL; X54530; G65159; -.
PIR; S12100; S12100.
HSSP; P02788; 1LGB.
PRINTS; [PR00422](#) TRANSFERRIN
PROSITE; PS00205; TRANSFERRIN_1; 2.
PROSITE; PS00206; TRANSFERRIN_2; 2.
PROSITE; PS00207; TRANSFERRIN_3; 2.
SIGNAL 1 22
CHAIN 23 701 TRANSFERRIN.
REPEAT 23 340 1.
DOMAIN 341 350 CONNECTING REGION.
REPEAT 351 381 2.

SWISS-PROT:P20233

PRINTS:PR00422

PubMed	Nucleotide	Protein	Genome
Search	Nucleotide	for	XLTRSFER
		Go	Clear
		Limits	Preview/Index
		History	Clipboard

Similarity searching

- If an identity search fails to find a match, **all is not lost**.
- The next step is to look for similar sequences
- It is recommended to perform similarity searches on peptides that are longer than ~ 30 residues (**the shorter the peptide, the greater the likelihood of finding chance matches that have no biological relevance**)

Similarity searching

- The most rapid and simple option is use BLAST (or FastA)
- We can use the sequence retrieved by from the primary database (OWL)

BLAST Output

- There are several important features to note in BLAST Output.
 - We are looking for matches that have high scores with correspondingly low probability values.
 - Low probability indicates that a match is unlikely to have arisen by chance.
 - The results show a cluster of high scores (with low probabilities) at the top of the list, indicating a likely relationship between the query and the family of sequences in the cluster.

Similarity searching

BLAST version 2.06

Search parameters	Output parameters
Expect: Default (10) or	Descriptions: Default (25)
Cutoff Score: Default or	Alignments: Default (10)
Matrix: BLOSUM62	Sort by: pvalue
Format: html ☒ or text ☐	
For DNA sequences only	
Strand: both	

Type of sequence:	protein ☒ or DNA ☐
Sequence submission:	protein code or accession ☐ or actual sequence ☒
Database:	OWL ☒ SWISS-PROT ☐ SP-TrEMBL ☐ SWISS-PROT + SP-TrEMBL ☐
Retrieve previous results:	search ID ☐

```
MDFSLRVALCLSMALCLAIQKEKQVRWCVKSNSSEL
KKCKDLVDTCNKKEIKLSCVEKSN
```

送出查詢

重設

BLAST Results

網址(D)  <http://bioinf.man.ac.uk/cgi-bin/dbbrowser/OWL/blast-cluster.cgi>

BLAST Results

Search ID no: 12995

BLASTP 2.0.9 [May-07-1999]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= USER_SEQUENCE
(60 letters)

Database: /nfs/seq/owl/owl.fasta
290,043 sequences; 93,009,033 total letters

Searching.....done

Sequences producing significant alignments:	Score (bits)	E Value
owl X54530 XLTRSFER XLTRSFER NID: g65158 - African clawed frog.	99	3e-21
owl P20233 TRFE_XENLA TRANSFERRIN PRECURSOR. - XENOPUS LAEVIS (...)	99	3e-21
owl P31226 SAX_RANCA SAXIPHILIN PRECURSOR (SAX). - RANA CATESBE...	45	5e-05
owl P56410 TRFE_ANAPL OVOTRANSFERRIN. - ANAS PLATYRHYNCHOS (DOM...	40	0.002
owl D89084 D89084 D89084 NID: g1694683 - Oncorhynchus kisutch c...	40	0.002
owl P02789 TRFE_CHICK OVOTRANSFERRIN PRECURSOR (CONALBUMIN) (AL...	39	0.004
owl X02009 GGCONR GGCONR NID: g63325 - chicken.	39	0.004
owl D64033 D64033 D64033 NID: g2143185 - Oryzias latipes DNA.	38	0.007
owl P80426 TRF1_SALSA SEROTRANSFERRIN I PRECURSOR (SIDEROPHILIN...	36	0.026
owl P80429 TRF2_SALSA SEROTRANSFERRIN II PRECURSOR (SIDEROPHILI...	36	0.026
owl X91908 GGEOS47GN GGEOS47GN NID: g1020103 - chicken.	35	0.058
owl P09571 TRFE_PIG SEROTRANSFERRIN (SIDEROPHILIN) (BETA-1-META...	33	0.22
owl P27425 TRFE_HORSE SEROTRANSFERRIN PRECURSOR (SIDEROPHILIN) ...	31	0.65
owl P08582 TRFM_HUMAN MELANOTRANSFERRIN PRECURSOR (MELANOMA-ASS...	30	1.5
owl J03298 MUSULT MUSULT NID: g202290 - Mouse (CD-1) uterine cD...	30	1.5
owl P19134 TRFE_RABIT SEROTRANSFERRIN PRECURSOR (SIDEROPHILIN) ...	30	1.9
owl AF031625 OCTF15 OCTF15 NID: g2736312 - Oryctolagus cuniculus.	30	1.9
owl AB010995 AB010995 AB010995 NID: g3786307 - Oryctolagus cuni...	29	2.5

Searching the Secondary databases

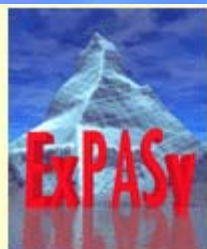
- Although a family of sequences was identified by the BLAST search, we can continue on and search the secondary database in order to discover if our query sequence contains any known characteristic conserved motifs.
- The first secondary database to consider is PROSITE.

Searching PROSITE

- At this example, the database code (TRFE_XENLA)
- Three regular expression patterns have been matched→
 - TRANSFERRIN_1
 - TRANSFERRIN_2
 - TRANSFERRIN_3

Y-x(0,1)-[VAS]-V-[IVAC]-[IVA]-[IVA]-[RKH]-[RKS]-[GDENSA]

Pattern of PS00205



ScanProsite

Enter an accession no.
(e.g., **P05130**), or ID
code (e.g., **KPCI_DROME**)

P20233

Or paste your sequence
here:

Exclude patterns with a
high probability of
occurrence ☐

RESET

STARTS THE SCAN

網址(D) <http://www.expasy.ch/cgi-bin/scanprosite/scanprosite?1>

[ExPASy Home page](#)

[Site Map](#)

[Search ExPASy](#)

ScanProsite - Protein against PROSITE

Scan of TRFE_XENLA ([P20233](#))

SEROTRANSFERRIN PRECURSOR.
Xenopus laevis (African clawed frog).

[1] [PDOC00004](#) [PS00004](#) **CAMP_PHOSPHO_SITE**

cAMP- and cGMP-dependent protein kinase phosphorylation site

Number of matches: 3

- | | | |
|---|---------|------|
| 1 | 118-121 | KKSS |
| 2 | 450-453 | KKGT |
| 3 | 526-529 | KKCS |

[2] [PDOC00005](#) [PS00005](#) **PKC_PHOSPHO_SITE**

Protein kinase C phosphorylation site

Number of matches: 8

- | | | |
|---|---------|-----|
| 1 | 4-6 | SLR |
| 2 | 45-47 | TCK |
| 3 | 120-122 | SSK |
| 4 | 166-168 | TWR |
| 5 | 249-251 | TRK |
| 6 | 252-254 | SIK |
| 7 | 522-524 | SER |
| 8 | 689-691 | TSR |

[3] [PDOC00006](#) [PS00006](#) **CK2_PHOSPHO_SITE**

Casein kinase II phosphorylation site

NiceSite View of PROSITE: [PS00205](#)

General information about the entry

Entry name	TRANSFERRIN_1
Accession number	PS00205
Entry type	PATTERN
Date	APR-1990 (CREATED); JUN-1994 (DATA UPDATE); JUL-1998 (INFO UPDATE).
PROSITE documentation	PDOC00182

Name and characterization of the entry

Description	Transferrins signature 1.
Pattern	Y-x(0,1)-[VAS]-V-[IVAC]-[IVA]-[IVA]-[RKH]-[RKS]-[GDENSA].

Numerical results

- SWISS-PROT release number: **38**, total number of sequence entries in that release: **80000**.
- Total number of hits in SWISS-PROT: **50 hits in 28 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **48 hits in 26 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- Number of false hits (on unrelated proteins): **2 hits in 2 different sequences**
- Number of known missed hits: **0**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **0**
- Precision (true hits / (true hits + false positives)): **96.00 %**
- Recall (true hits / (true hits + false negatives)): **100.00 %**

Sn; SP|P20233|TRFE_XENLA|TRANSFERRIN PRECURSOR.

1TBS

1TBH [TRANSFERRIN](#) 1.3e-45

1TBF

2TBS

2TBT	FingerPrint	No.Motifs	SumId	AveId	ProfScore	Ppvalue	Evalue	GraphScan
2TBH	TRANSFERRIN	5 of 5	3.5e+02	69	3868	1.4e-53	1.3e-45	IIIII
2TBN	GATAZNFINGER	2 of 2	58.38	29.19	485	0.00033	3e+04	ii
2TBN	GPCRMGR	2 of 12	54.18	27.09	373	0.0012	1e+05	I.....i.
2TBN	5HT1DRECEPTR	2 of 5	62.03	31.02	409	0.0023	2e+05	i..I.
2TBN	PHYTOCHROME	2 of 10	60.01	30.01	420	0.0038	3.4e+05	.i.I.....
2TBN	5HT5BRECEPTR	2 of 6	46.67	23.33	375	0.0042	3.8e+05	..i.i.
2TBN	LYMPHSPCIFIC	2 of 10	61.43	30.72	396	0.0058	5.2e+05	..I....i..
2TBN	ADENOSINEAIR	2 of 5	74.18	37.09	437	0.0071	6.4e+05	..I.I

2TBF

3TBS

3TBT	MotifName	No.Mots	IdScore	PfScore	Pvalue	Sequence	Len	low	pos	high
3TBH	TRANSFERRIN	1 of 5	68.30	878	1.69e-13	YYNYAGAFKCLQDDQGDVAFVK	22	149	208	352
3TBH	TRANSFERRIN	2 of 5	74.73	803	3.24e-13	ADAVTLDDGGYMYTAGLCGLV	20	338	398	537
3TBH	TRANSFERRIN	3 of 5	72.46	901	7.70e-12	LRGVKTCHTAVGRTAGWNIPV	21	396	460	602
3TBH	TRANSFERRIN	4 of 5	72.69	768	1.66e-10	WAKDLKSEDFELLCPDGSR	19	507	572	713
3TBH	TRANSFERRIN	5 of 5	58.98	518	2.03e-07	CNLAEVPAHAVVTLPDK	17	535	599	740
3TBB										
3TBN	GATAZNFINGER	1 of 2	27.50	294	3.85e-03	EAQKTQECKLFSSPVWGR	18	1	287	739
3TBN	GATAZNFINGER	2 of 2	30.89	191	8.59e-02	DGGYMYTAGLCGLVPVMG	18	19	404	757
3TBB										
3TBN	GPCRMGR	1 of 12	36.38	202	2.00e-02	GGDVYKGSLLQPYN	13	29	80	84
3TBN	GPCRMGR	11 of 12	17.79	171	5.82e-02	STSRLLAACTFHAA#####	24	734	688	844
3TBB										
3TBN	5HT1DRECEPTR	1 of 5	26.32	200	6.17e-02	SQLCAGIKEHKCSRSNNEP	19	5	189	5
3TBN	5HT1DRECEPTR	4 of 5	35.71	209	3.68e-02	TAGLCGLVPVMGEY	14	247	410	250
3TBB										
3TBN	PHYTOCHROME	2 of 10	29.29	239	4.09e-02	EPYYNYAGAFKCLQDDQGDV	20	143	206	274
3TBN	PHYTOCHROME	4 of 10	30.73	181	9.40e-02	IVQILKGDADAVTLDDGGMYT	21	340	390	472
3TBB										
3TBN	5HT5BRECEPTR	3 of 6	26.67	195	5.24e-02	DGGDVYKGSLLQPYNL	15	36	79	36
3TBN	5HT5BRECEPTR	5 of 6	20.00	180	7.99e-02	DSAVSLIPLPSSMDS	15	246	310	246
3TBB										
3TBN	LYMPHSPCIFIC	3 of 10	35.29	193	8.88e-02	LLLERKLLKWAGPDSET	17	130	150	133
3TBN	LYMPHSPCIFIC	8 of 10	26.14	203	6.50e-02	QEAQKTQECKLFSSPVWGRDLI	22	266	286	275

Searching Pfam

- Another important resource to search is the Pfam collection of Hidden Markov Models.
- The sequence must be in FastA format.

Searching PRINTS

- The Output of PRINTS is divided into distinct sections:
 - The program offers an intelligent ‘guess’ based on the occurrence of the highest-scoring complete or partial fingerprint match-it then provides an expanded table that shows the top 10 best-scoring matches.
 - If the guess is considered either to be wrong or to have missed something, the remaining sections of output provide more of the row data, again allowing the user to search for anything that might have been missed.

Searching IDENTIFY

- The output of IDENTIFY is normally given at several stringency levels

When to believe a result

- The real need for such a strategy arise when query sequences do not have obvious homologues in the database, i.e., do not return positive identifications following sequence identity or similarity searches of the primary database.
- Under these circumstances, composing a diagnostic picture from searches of a range of secondary databases is essential. If two or more of these make twilight matches, this builds confidence in making a diagnosis.