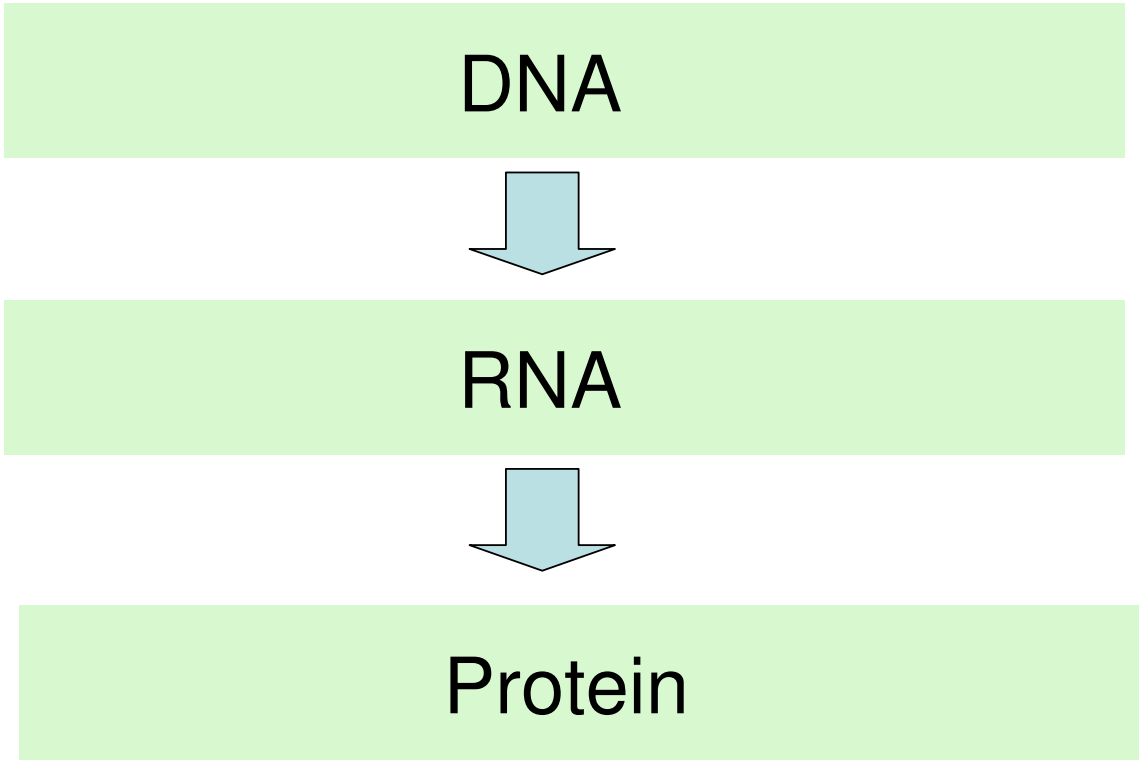
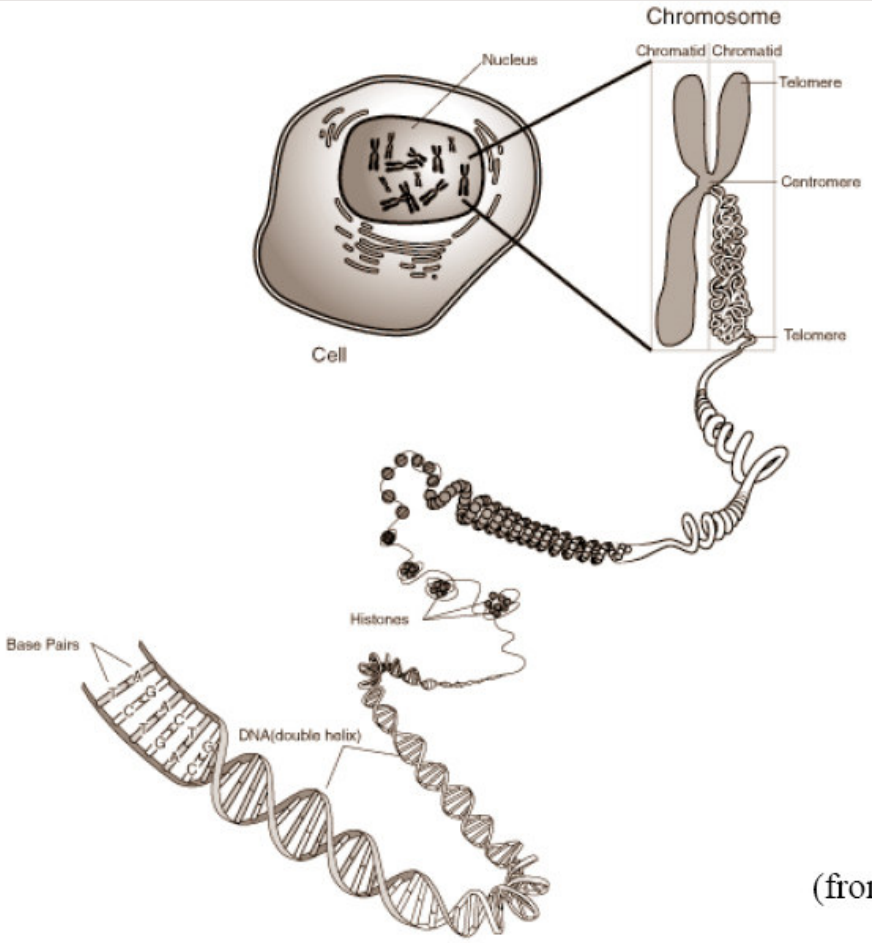
Central Dogma of Molecular Biology
(생화학의 중심원리)



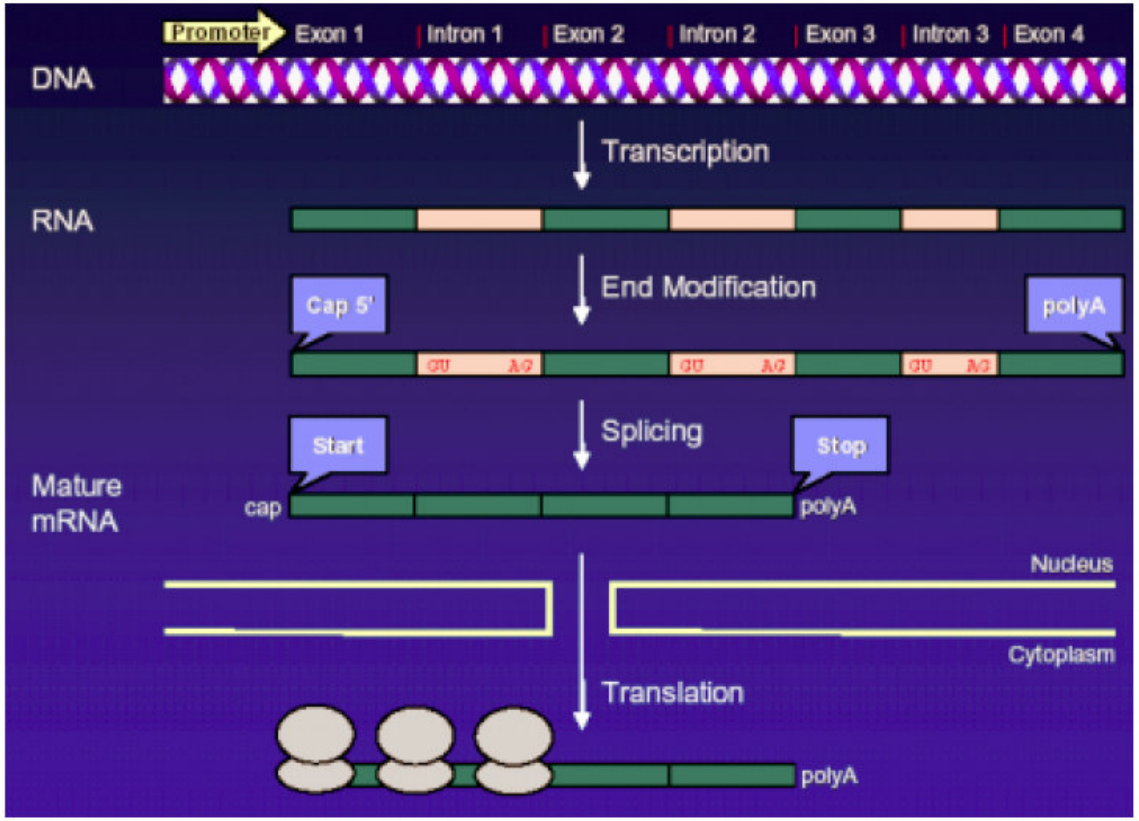
Chromosome, Genome, and Gene



- Prokaryotes vs. eukaryotes
- Human
  - 22 pairs of chromosomes + sex chromosome
  - 3.2 billion bp
  - ~35,000 genes
  - 90-95% junk DNA

(from NIH genetic illustrations gallery)

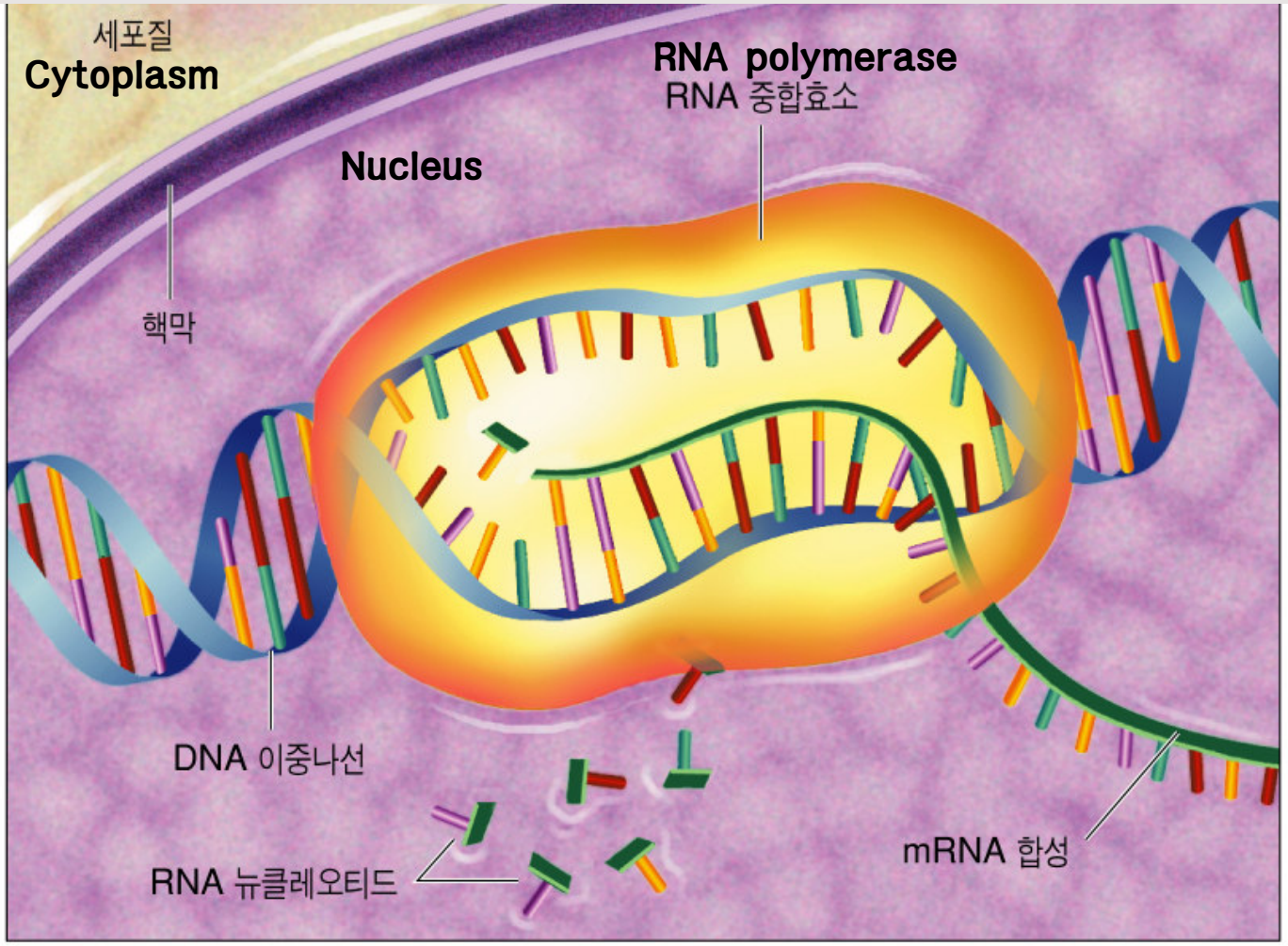
DNA to RNA (Transcription)



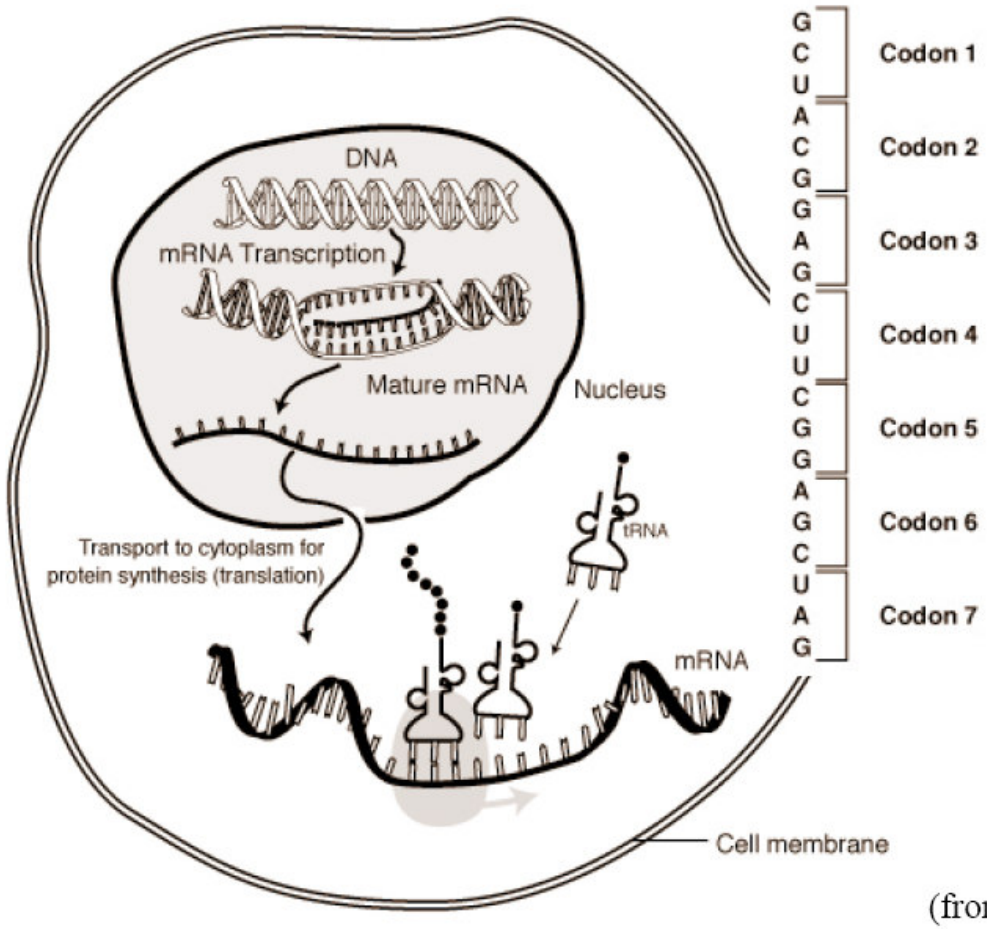
- RNA polymerase
- Promoter
- Exon-intron
  - no intron in prokaryotes
  - Avg 200 bp vs. 10 Kbp
- PolyA signal
- Splicing
- Alternative splicing

(Picture from A. Baxevanis's lecture)

DNA to RNA (Transcription)



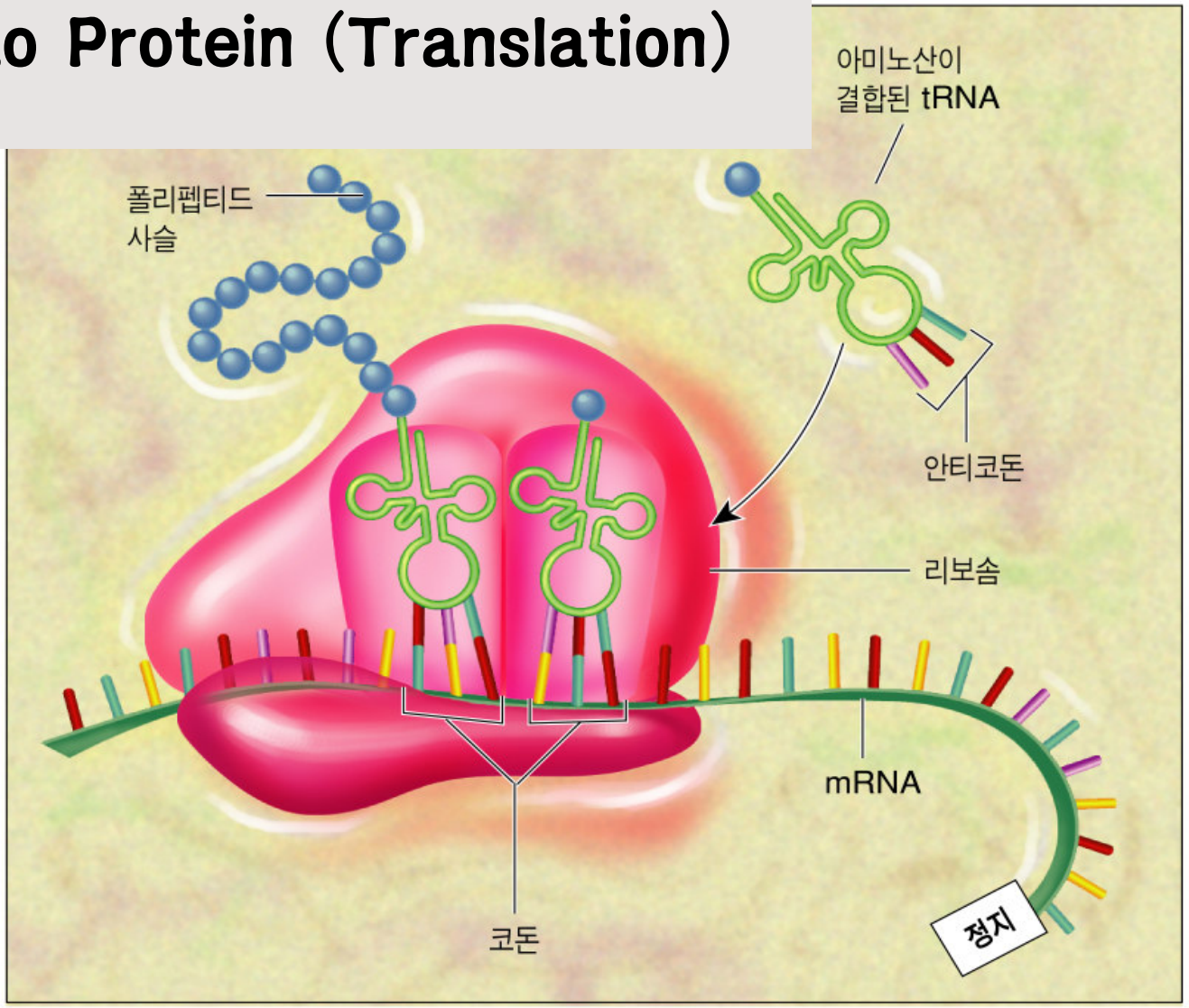
RNA to Protein (Translation)



- Ribosome
- Codon triplet
  - Genetic codes
  - tRNA
  - anticodon
- Coding region vs. noncoding region
  - Regulons
  - 5′ UTR, 3′ UTR

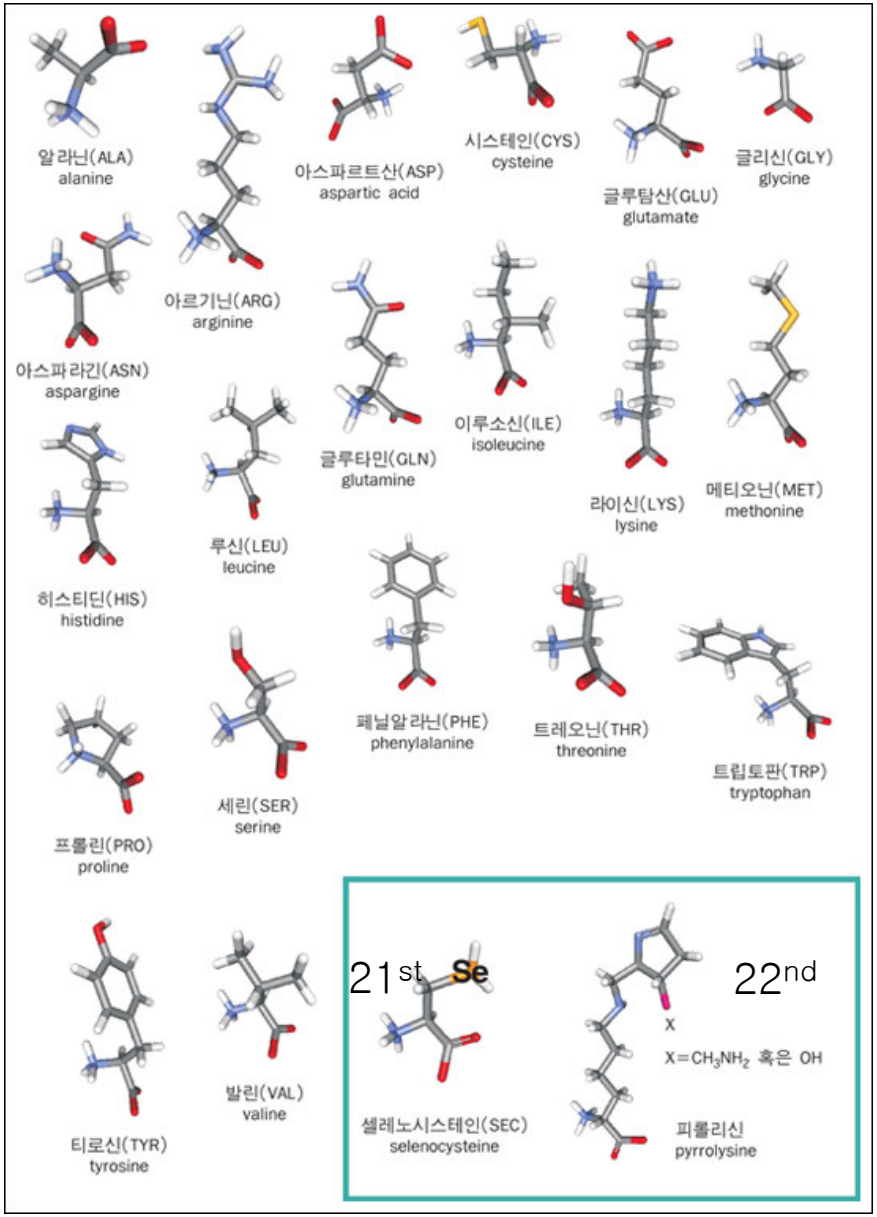
(from NIH genetic illustrations gallery)

RNA to Protein (Translation)



경상대학교 생화학과 생물정보학연구실
Bioinformatics Lab http://bio.gsnu.ac.kr

Molecular Structure of Amino Acids



Basic Properties of Amino Acids

도표 3-2. 20가지 기본 아미노산에 대한 특성

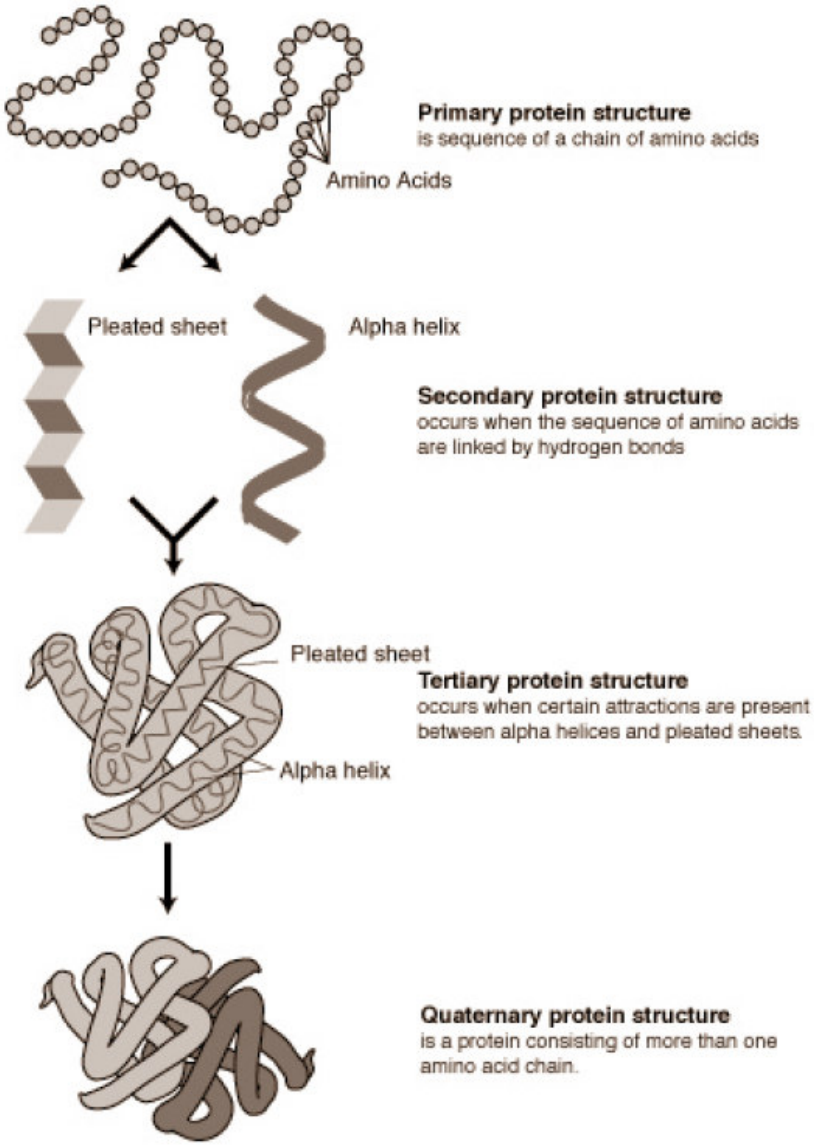
아미노산	3문자 코드	1문자 코드	질량	표면적	부피	pka	등전위점값(pI)	용해도	밀도
알라닌	ALA	A	71.09	115	88.6	–	6.107	16.65	1.401
아르기닌	ARG	R	156.19	225	173.4	~12	10.76	15 1.1	
아스파르트산	ASP	D	114.11	150	111.1	4.5	2.98	0.778	1.66
아스파라긴	ASN	N	115.09	160	114.1	–	–	3.53	1.54
시스테인	CYS	C	103.15	135	108.5	9.1~9.5	5.02	매우 높음	–
글루탐산	GLU	E	129.12	190	138.4	4.6	3.08	0.864	1.46
글루타민	GLN	Q	128.14	180	143.8	–	–	2.5	–
글리신	GLY	G	57.05	75	60.1	–	6.604	24.99	1.607
히스티딘	HIS	H	137.14	195	153.2	6.2	7.64	4.19	–
이소루신	ILE	I	113.16	175	166.7	–	6.038	4.117	–
루신	LEU	L	113.16	170	166.7	–	6.036	2.426	1.191
라이신	LYS	K	128.17	200	168.6	10.4	9.47	매우 높음	–
메티오닌	MET	M	131.19	185	162.9	–	5.74	3.381	1.34
페닐알라닌	PHE	F	147.18	210	189.8	–	5.91	2.965	–
프롤린	PRO	P	97.12	145	112.7	–	6.3	162.3	–
세린	SER	S	87.08	115	89	–	5.68	5.023	1.537
트레오닌	THR	T	101.11	140	116.1	–	–	매우 높음	–
트립토판	TRP	W	186.12	255	227.8	–	5.88	1.136	–
티로신	TYR	Y	163.18	230	193.6	9.7	5.63	0.0453	1.456
발린	VAL	V	99.14	155	140	–	6.002	8.85	1.23

자료 출처 : Information from NIST Chemistry Webbook(2003)

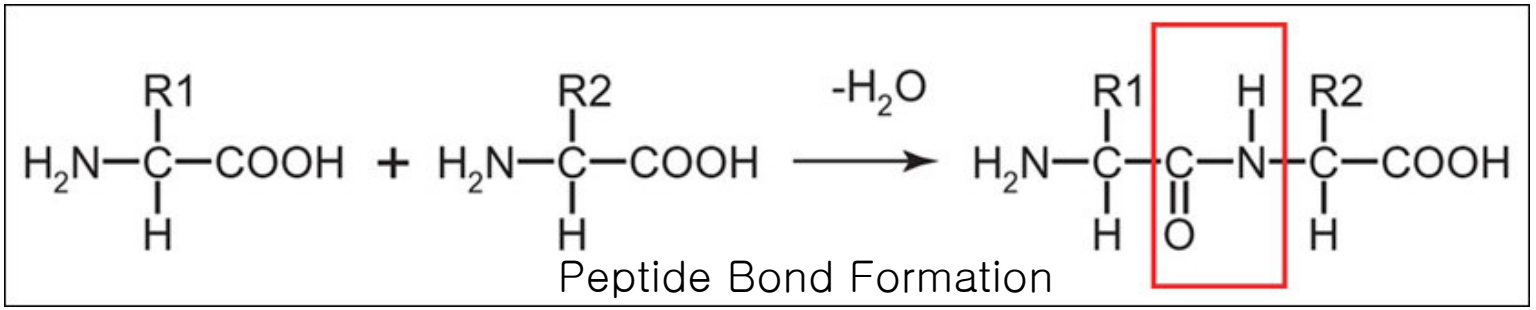
[a] 질량[dalton], 표면적[Å^2], 부피[Å^3], pka[잔기], pI[25℃], 용해도[g/100g, 25℃], 밀도[결정밀도, g/ml]

Proteins

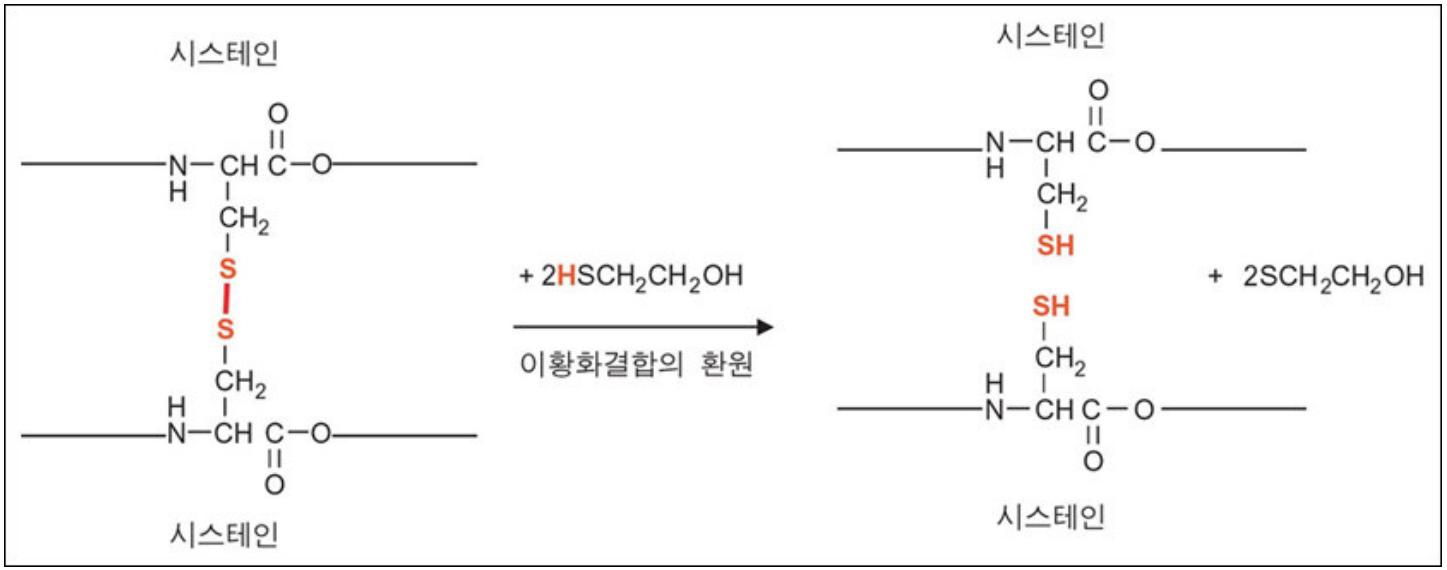
- Structures
  - Primary
  - Secondary
    - α-helices, β-sheets
  - Tertiary
    - Domains
  - Quaternary
- Structure prediction
  - Homology
  - Threading
  - Ab initio folding



AA-AA Binding using Polypeptide bond



Disulfide Bond (Bridge) by Cysteine, Cys



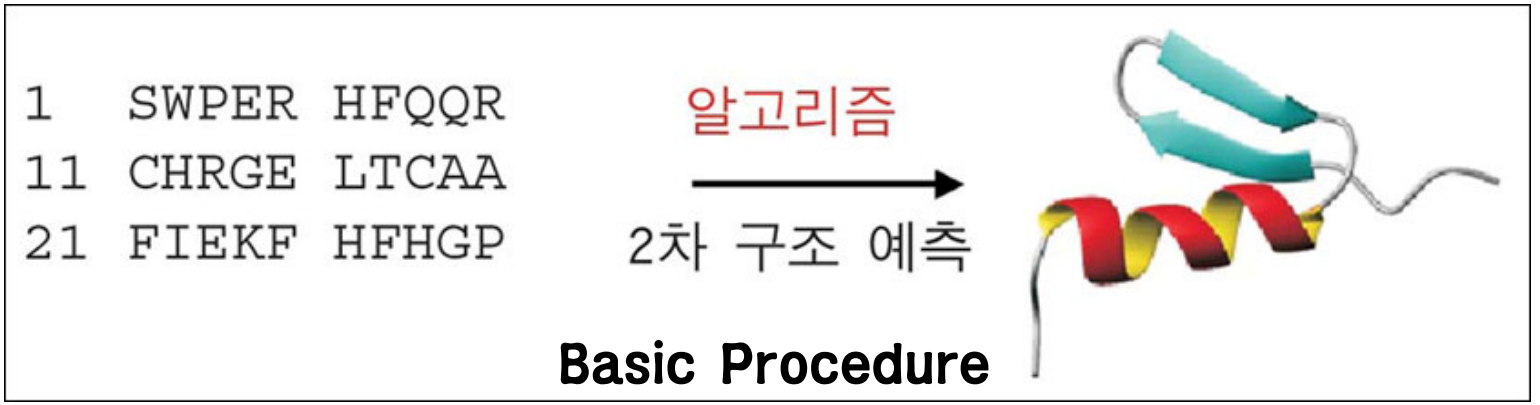
Amino Acid Codon

Codon: 3 DNA bases (A, T, C, G) to recognize amino acids

4 X 4 X 4 = 64 **개 가능**
그러나 중복코돈이 많음

첫 번째 핵산	두 번째 핵산				세 번째 핵산
	U	C	A	G	
U	UUU 페닐알라닌 (Phe)	UCU 세린 (Ser)	UAU 티로신 (Tyr)	UGU 시스테인 (Cys)	U
	UUC 페닐알라닌 (Phe)	UCC 세린 (Ser)	UAC 티로신 (Tyr)	UGC 시스테인 (Cys)	C
	UUA 루신 (Leu)	UCA 세린 (Ser)	UAA 정지, 글루타민[1]	UGA 정지, 트립토판[2, 3] 시스테인[4] 셀레노시스테인[5]	A
	UUG 루신 (Leu)	UCG 세린 (Ser)	UAG 정지, 글루타민[1]	UGG 트립토판 (Trp)	G
C	CUU 루신 (Leu)	CCU 프롤린 (Pro)	CAU 히스티딘 (His)	CGU 아르기닌 (Arg)	U
	CUC 루신 (Leu)	CCC 프롤린 (Pro)	CAC 히스티딘 (His)	CGC 아르기닌 (Arg)	C
	CUA 루신 (Leu)	CCA 프롤린 (Pro)	CAA 글루타민 (Gln)	CGA 아르기닌 (Arg)	A
	CUG 루신 (Leu), 세린 (Ser)[6]	CCG 프롤린 (Pro)	CAG 글루타민 (Gln)	CGG 아르기닌 (Arg)	G
A	AUU 이소루신 (Ile)	ACU 트레오닌 (Thr)	AAU 아스파라긴 (Asn)	AGU 세린 (Ser)	U
	AUC 이소루신 (Ile)	ACC 트레오닌 (Thr)	AAC 아스파라긴 (Asn)	AGC 세린 (Ser)	C
	AUA 이소루신 (Ile)	ACA 트레오닌 (Thr)	AAA 라이신 (Lys)	AGA 아르기닌 (Arg)	A
	AUG 메티오닌 (Met) 혹은 시작	ACG 트레오닌 (Thr)	AAG 라이신 (Lys)	AGG 아르기닌 (Arg)	G
G	GUU 발린 (Val)	GCU 알라닌 (Ala)	GAU 아스파르트산 (Asp)	GGU 글리신 (Gly)	U
	GUC 발린 (Val)	GCC 알리닌 (Ala)	GAC 아스파르트산 (Asp)	GGC 글리신 (Gly)	C
	GUA 발린 (Val)	GCC 알라닌 (Ala)	GAA 글루탐산 (Glu)	GGA 글리신 (Gly)	A
	GUG 발린 (Val)	GCG 알라닌 (Ala)	GAG 글루탐산 (Glu)	GGG 글리신 (Gly)	G

Protein 2차구조 예측: Methods



A. 이미알려진 단백질의 구조와 정보 활용
B. 물리-화학적 특성활용

- C-F (Chou-Fasman) Method : Q_3=50~55%
- GOR (Garnier-Osguthorpe-Robson) Method: Q_3=~63%
- Lim's Method: Q_3=~65%
- Neural Network Method: Q_3=~75%

경상대학교 생화학과 생물정보학연구실
Bioinformatics Lab http://bio.gsnu.ac.kr

Protein 2차구조예측: C-F Method

1978년 Chou와 Fasman에 의하여 제안됨.

Basic Procedure

Step1. 각 아미노산의 발생빈도 분석후 도표화. (도표이용)

Step2. 주어진 서열의 helix 와 sheet 가능성을 판별하고 양방향으로 확장.

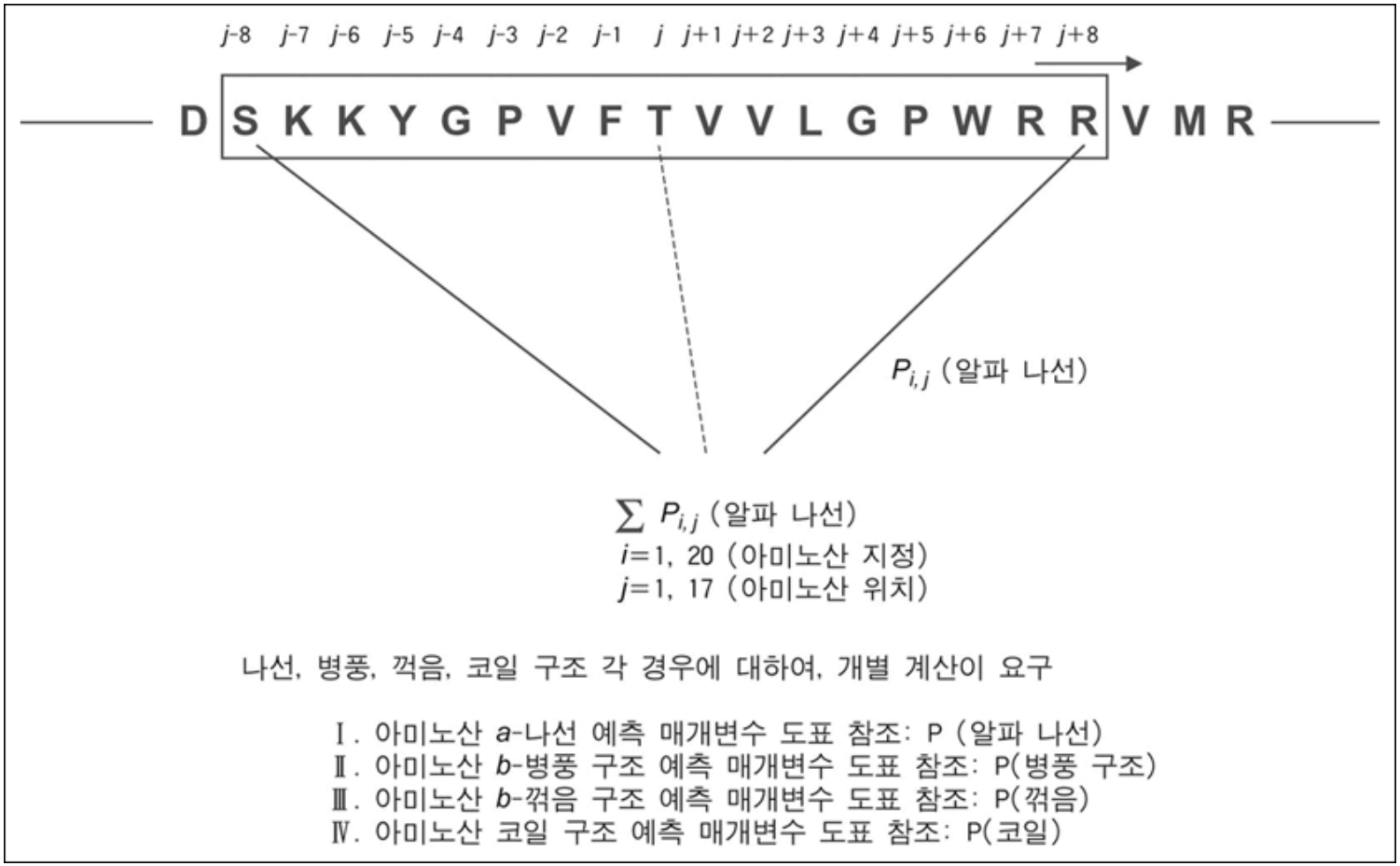
Step3. 같은 위치에 두구조가능성이 겹칠경우 P(helix) 와 P(sheet) 의 평균치를 비교하여 높은것으로 할당.

Step4. Turn의 경우 연속되는 4개의 아미노산을 중심으로 2단계 개별 확률치로 수치화.

경상대학교 생화학과 생물정보학연구실
Bioinformatics Lab http://bio.gsnu.ac.kr

Protein 2차구조예측: GOR Method

1978년 Garnier-Osguthorpe-Robson 이 개발

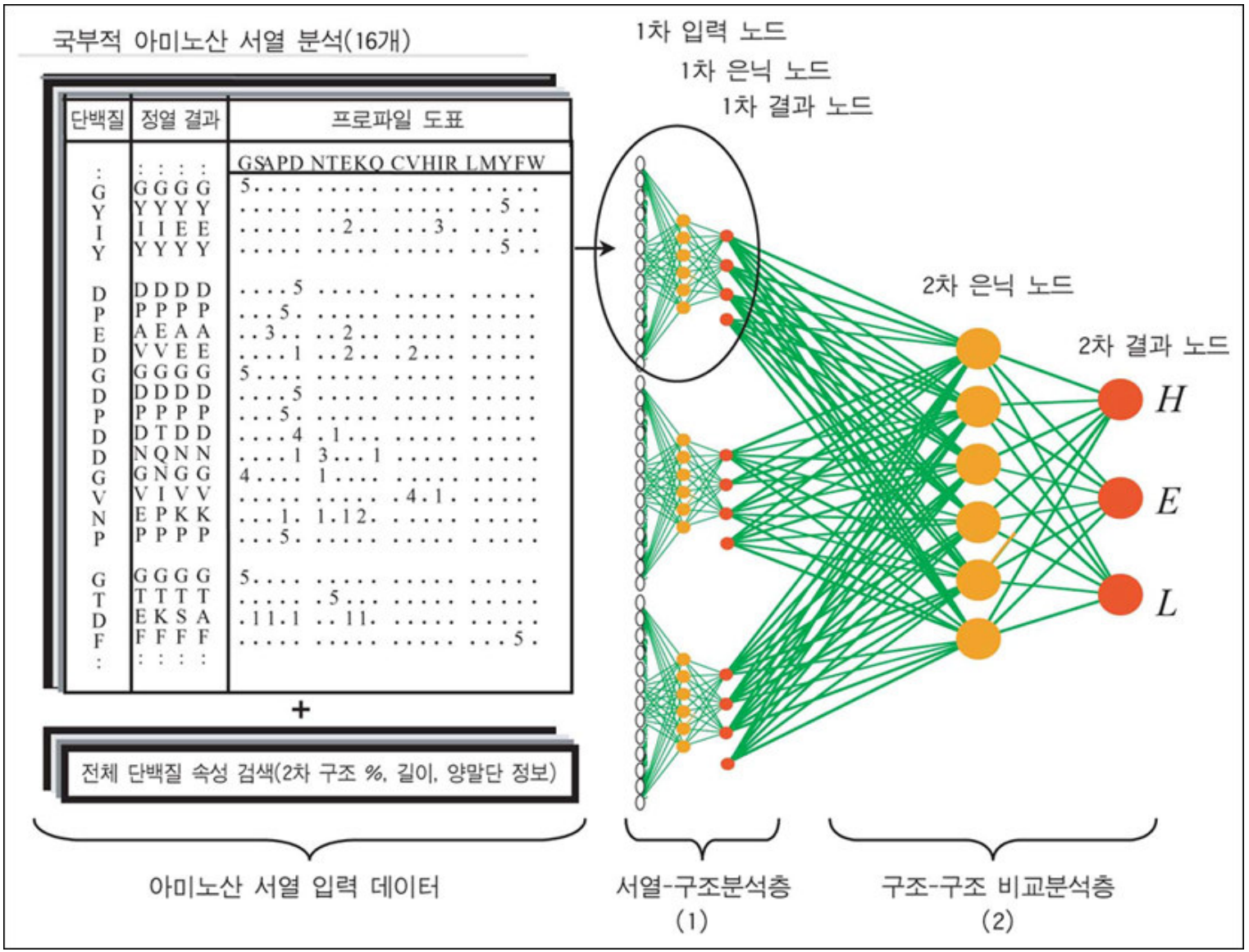


나선, 병풍, 꺾음, 코일 구조 각 경우에 대하여, 개별 계산이 요구

Ⅰ. 아미노산 *a*-나선 예측 매개변수 도표 참조: P (알파 나선)
Ⅱ. 아미노산 *b*-병풍 구조 예측 매개변수 도표 참조: P(병풍 구조)
Ⅲ. 아미노산 *b*-꺾음 구조 예측 매개변수 도표 참조: P(꺾음)
Ⅳ. 아미노산 코일 구조 예측 매개변수 도표 참조: P(코일)

경상대학교 생화학과 생물정보학연구실
Bioinformatics Lab http://bio.gsnu.ac.kr

Protein 2차구조예측: Neural Network Method



Retrieval, Sequence Search & Classification Methods

- Retrieve protein info by text / UID
- Sequence Similarity Search
  - BLAST, FASTA, Dynamic Programming
- Family Classification
  - Patterns, Profiles, Hidden Markov Models, Sequence Alignments, Neural Networks
- Integrated Search and Classification System

Sequence Similarity Search

- Based on Pair-Wise Comparisons
- Dynamic Programming Algorithms
  - Global Similarity: Needleman-Wunch
  - Local Similarity: Smith-Waterman
- Heuristic Algorithms
  - FASTA: Based on K-Tuples (2-Amino Acid)
  - BLAST: Triples of Conserved Amino Acids
  - Gapped-BLAST: Allow Gaps in Segment Pairs
  - PHI-BLAST: Pattern-Hit Initiated Search
  - PSI-BLAST: Position-Specific Iterated Search

BLAST

BALST (Basic Local Alignment Search Tool)

- Extremely fast
- Robust
- Most frequently used

It finds very short segment pairs (“seeds”) between the query and the database sequence

These seeds are then extended in both directions until the maximum possible score for extensions of this particular seed is reached

BLAST Search

- **From BLAST Search Interface**
- **Table-Format Result with BLAST Output and SSEARCH (Smith-Waterman) Pair-Wise Alignment**

Query sequence: (NF00682686, length=170, Search iProClass, e-value < 0.0001, filter=T)
>NF00682686 hypothetical protein F24I3.170 [Arabidopsis thaliana]
MDAKIGQFFDSVGTFFSGSDKIPWCDGDVIAGCEREVREATDSGTEDLKKECLMRLSWAL
VHSRQTEDVQRGIAMLEASLESSAPPLEDREKLYLLAVGYYRSGNYSRSRQLVDRCIEMQ
ADWRQALVLKKTIEDKITKDGVIGIGITATAFGAVGLIAGGIVAAMSRKK

9 match(es) shown in the following table: HELP

For sequence analyses, pick a method (radiobutton) below, select a sequence(s) (checkbox) in *Protein ID* column, and GO.
BLAST FASTA HMM Search Pattern Match Multiple Alignment Domain Display GO

Protein ID (check all)	Protein Name	Organism	Taxon Group	Superfamily	Family	e-value	Length	Ov.lap	% idn	Query Sequence
NREF:NF00682686 iProClass:Q9M1J1 PIR-PSD:T47769 SP/TR:Q9M1J1	hypothetical protein F24I3.170	Arabidopsis thaliana(mouse-ear cress)	Euk/Plant	SF026168	FAM0097587	8e-85	170	170	100	
NREF:NF00681186 iProClass:Q94CK3 SP/TR:Q94CK3	Hypothetical protein	Arabidopsis thaliana(mouse-ear cress)	Euk/Plant			9e-46	167	168	57	
NREF:NF00950927 iProClass:Q8RZ98 SP/TR:Q8RZ98	B1147A04.28 protein	Oryza sativa(rice)	Euk/Plant			7e-44	199	169	60	
NREF:NF01489251 iProClass:Q7S8M1 SP/TR:Q7S8M1	Hypothetical protein	Neurospora crassa	Euk/Fungi-Metazoa			2e-09	153	114	35	

BLAST/SSEARCH Results

NREF:NF00950927 iProClass:Q8RZ98 SP/TR:Q8RZ98	B1147A04.28 protein	Oryza sativa(rice)	Euk/Plant			7e-44	199	169	60

```
Smith-Waterman score: 605;  60.355% identity in 169 aa overlap

                                        10        20        30
NF0068                               MDAKIGQFFDSVGTFFSGSDKIPWCDGDVI
                                     :.::::..  ..:.::::.:..:::  :.:
NF0095 SIATWPEILLHRLRAKPSRFLPHRSRRSAAMEAKIGRLVGAIGAFFSGGDNVPWCGRDII
          20        30        40        50        60        70

               40        50        60        70        80        90
NF0068 AGCEREVREATDSGTEDLKKECLMRLSWALVHSRQTEDVQRGIAMLEASLESSAPPLEDR
       :: :::: ::.   ::. :.  .:::::::::::.:.::.:::::.::: .:  ::: :
NF0095 AGVEREVAEAA---TEEHKNVSIMRLSWALVHSRNTDDVNRGIAMLQASLGGSKSPLEAR
          80            90        100       110       120       130

              100       110       120       130       140       150
NF0068 EKLYLLAVGYYRSGNYSRSRQLVDRCIEMQADWRQALVLKKTIEDKITKDGVIGIGITAT
       ::::::::::.::.         :  ::..:  : ::: ::::.::::.::::::::..:
NF0095 EKLYLLAVGHYRN---------VATCIQIQPGWGQALSLKKTVEDKIAKDGVIGIGIATT
            140                 150       160       170       180

              160       170
NF0068 AFGAVGLIAGGIVAAMSRKK
       :   :::..: :.::..::.
NF0095 A---VGLLVG-IAAAVARKN
              190
```

SSEARCH Alignment

BLAST Alignment

```
 Score =  177 bits (448), Expect = 7e-44
 Identities = 93/151 (61%), Positives = 108/151 (70%), Gaps = 12/151 (7%)

Query: 1   MDAKIGQFFDSVGTFFSGSDKIPWCDGDVIAGCEREVREATDSGTEDLKKECLMRLSWAL 60
           M+AKIG+   ++G FFSG D +PWC  D+IAG EREV EA    TE+ K   +MRLSWAL
Sbjct: 46  MEAKIGRLVGAIGAFFSGGDNVPWCGRDIIAGVEREVAEAA---TEEHKNVSIMRLSWAL 102

Query: 61  VHSRQTEDVQRGIAMLEASLESSAPPLEDREKLYLLAVGYYRSGNYSRSRQLVDRCIEMQ 120
           VHSR T+DV RGIAML+ASL  S  PLE REKLYLLAVG+YR+         V  CI++Q
Sbjct: 103 VHSRNTDDVNRGIAMLQASLGGSKSPLEAREKLYLLAVGHYRN---------VATCIQIQ 153

Query: 121 ADWRQALVLKKTIEDKITKDGVIGIGITATA 151
             W QAL LKKT+EDKI KDGVIGIGI  TA
Sbjct: 154 PGWGQALSLKKTVEDKIAKDGVIGIGIATTA 184
```

Remote Homology Detection

- Psi-BLAST/RPS-BLAST
- HMMs: HMMER, SAM
- Domain databases
- Fold recognition approaches (Meta Servers)

Protein Domain Databases

Selection

- PFAM
  - http://pfam.wustl.edu/
- PROSITE
  - http://us.expasy.org/prosite/
- ProDom
  - http://prodes.toulouse.inra.fr/prodom/2002.1/html/home.php
- InterPro
  - http://www.ebi.ac.uk/interpro/

Family Classification Methods

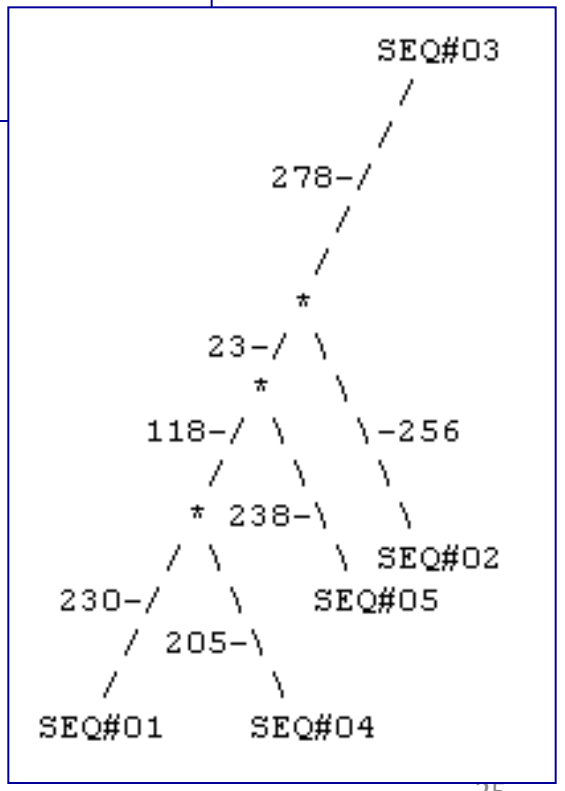
- Based on Family Information
- ClustalW Multiple Sequence Alignment
- ProSite Pattern Search
- Profile Search
- Hidden Markov Models (HMMs)
- Neural Networks
- Integrated Analysis

Multiple Sequence Alignment

```
NF00682686      FSGS-----DKIPWCDGDVIAGCEREVREATDSGTEDLKKECLMRLSWALVHSRQTEDVQ
NF00950927      FSGG-----DNVPWCGRDIIAGVEREVAEA---ATEEHKNVSIMRLSWALVHSRNTDDVN
NF00681186      FSGAASASADEFPLCDSDIISGCEKELAEA---QDEGRKKECIMRLSWALVHSKMPSDIQ
NF00535471      ----------------VEDLKNFERKFQSE--QAAGSVSKSTQFEYAWCLVRSKYNEDIR
                              : : . *::. .         .:   :. :*.**:*:  .*:.

NF00682686      RGIAMLEASLESSAPPLEDREKLYLLAVGYYRSGNYSRSRQLVDRCIEMQADWRQALVLK
NF00950927      RGIAMLQASLGGSKSPLEAREKLYLLAVGHYR--N-------VATCIQIQPGWGQALSLK
NF00681186      RGIAMLEALVVNDTSAMKLREKLYLLALGYYRSGDFSRSRDCIERCLEVEPESGQAQALK
NF00535471      RGIVLLEELLPKG-SKEEQRDYVFYLAVGNYRLKEYEKALKYVRGLLQTEPQNNQAKELE
                ***.:*:  :  . .  : *: :: **:* **  :      :   :: :.   **  *:

NF00682686      KTIEDKITKDGVIGIGITATAFGAVGLIAGGIVAAMSRKK-
NF00950927      KTVEDKIAKDGVIGIGIATT---AVGLLVG-IAAAVARKN-
NF00681186      KAIEDRIVKDGVIGVGIAVT---AVGVVAG-IAAAILRS--
NF00535471      RLIDKAMKKDGLVGMAIVGGMALGVAGLAGLIGLAVSKSKS
                : ::. : ***::*:.*.     .*. :.* *  *: :.
```



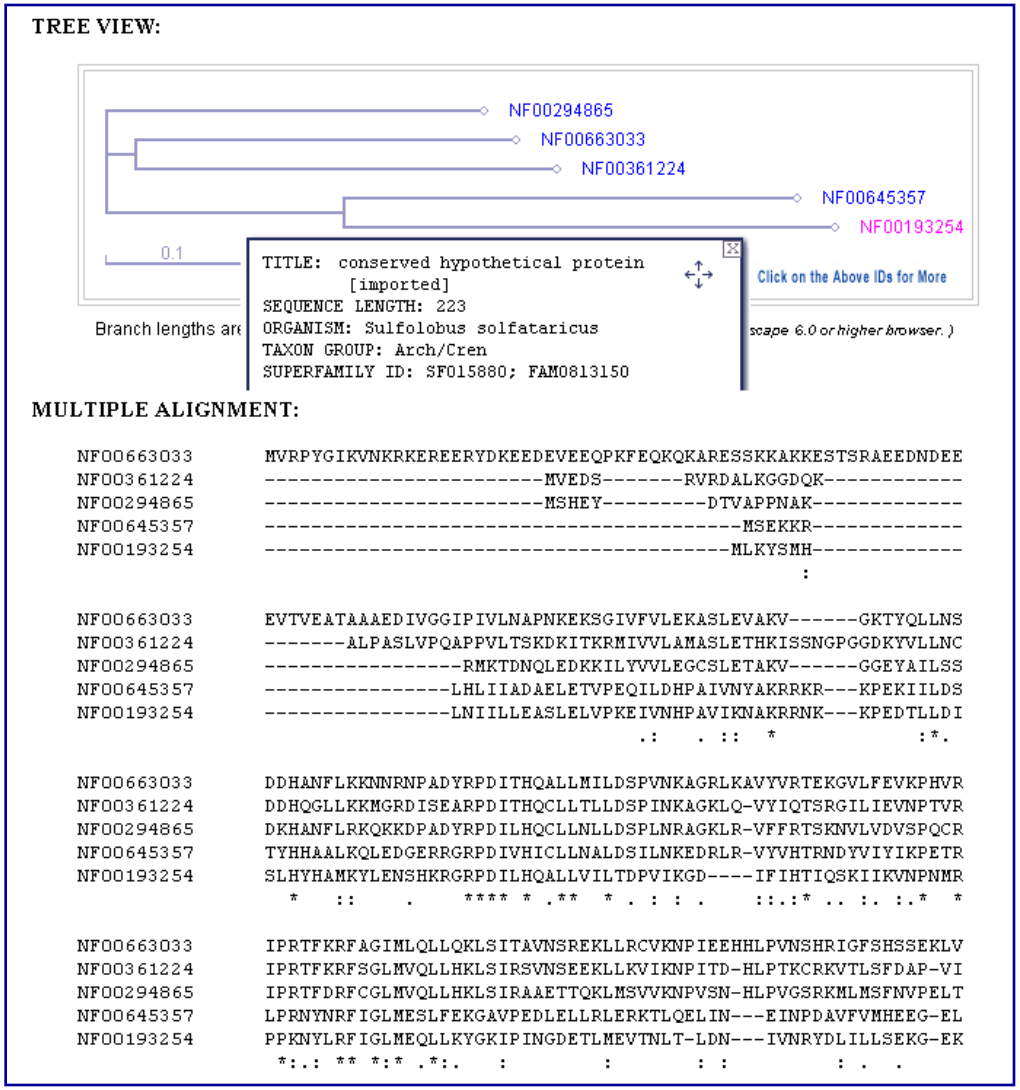
- ClustalW
- Progressive Pairwise Approach
  - Base on Exhaustive Pairwise Alignments
- Neighbor Joining
  - Joining Order Corresponding to a Tree
- Alignment Varies
  - Dependent on Joining Order

How do you build a tree?

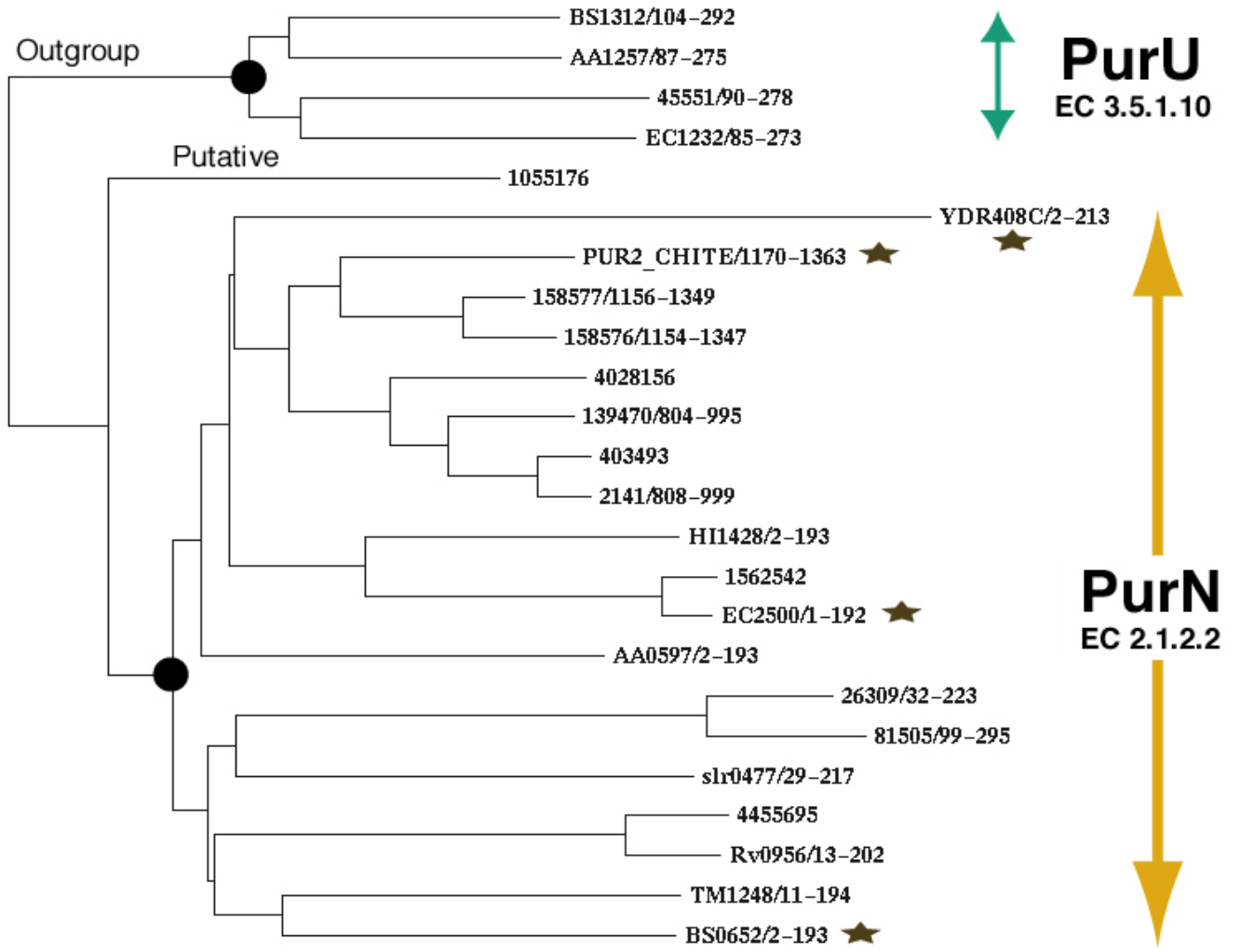
- Pick sequences to align
- Align them
- Verify the alignment
- Keep the parts that are aligned correctly
- Build and evaluate a phylogenetic tree

Multiple Alignment and Tree

- **From Text/Sequence Search Result or ClustalW Alignment Interface**



PurN Family Tree with Outgroup PurU



★ Function Experimentally Determined

Phylogenetic Tree by Neighbor Joining

Motif Patterns (Regular Expressions)

- Signature Patterns for Functional Motifs

```
PCM_AC  PCM00836
PCM_ID  ALADH_PNT_1; MOTIF
PS_DE   Alanine dehydrogenase & pyridine nucleotide transhydrogenase signature 1
PS_PA   G-[LIVM]-P-x-E-x(3)-N-E-x(1,3)-R-V-A-x-[ST]-P-x-[GST]-V-x(2)-L-x-[KRH]-x-G.
PROSITE PS00836; PDOC00654
LENGTH  Conserve = 16aa; Maximum = 29aa; Minimum = 27aa
COUNT   PST = 5 (5); PSN = 2 (2); PCT = 2 (2); PCN = 3 (3);

NNTM_BOVIN+DEBOXM     PST      60  GVPKEIFQNEK--RVALSPAGVQALVKQG
+G02257               PCT      60  GVPKEIFQNEK--RVALSPAGVQNLVKQG
DHA_BACSH+A34261      PST       4  GIPKEIKNNEN--RVAMTPAGVVSLTHAG
DHA_BACST+B34261      PST       4  GIPKEIKNNEN--RVAITPAGVMTLVKAG
DHA_MYCTU+A43830      PST       4  GIPTETKNNEFQFRVAITPAGVAELTRRG
PNTA_ECOLI+DEECXA     PST       4  GIPRERLTNET--RVAATPKTVEQLLKLG
                                   *:* *   **   *** :*  *  * : *
DHA_BACSU+A49337      PSN       4  GVPKEIKNNEN--RVALTPGGVSQLISNG
PNTA_HAEIN+E64119     PSN       4  GVPRELLENES--RVAATPKTVQQILKLG
+S74638               PCN       4  GVPKEIKDQEF--RVGLTPSSVRALLSQG
+S77433               PCN      23  GVPRESFDQEC--RVAMTPDTAQKLQKLG
+F64694               PCn       4  GLVKESMDLES--RVALVPDDVALIVQKG
                                   *:  *   *   **.  *  .  :   *
```

ProClass Motif Alignments

	Member	Non-Member
Predicted	True Positive (“T”)	False Positive (“F”)
Not Predicted	False Negative (“N”)	True Negative

PIR Pattern Search

- From Text/Sequence Search Result or Pattern Search Interface
- One Query Sequence Against PROSITE Pattern Database
- One Query Pattern (PROSITE or User-Defined) Against Sequence DB

Pattern Search Result (I)

- **One Query Sequence Against PROSITE Pattern Database**

Your input sequence B37245 matches 8 Prosite pattern(s).

1> PS00001 ASN GLYCOSYLATION; PATTERN.
2> PS00005 PKC PHOSPHO SITE; PATTERN.
3> PS00006 CK2 PHOSPHO SITE; PATTERN.
4> PS00008 MYRISTYL; PATTERN.
5> PS00009 AMIDATION; PATTERN.
6> PS00464 RIBOSOMAL L22; PATTERN.
7> PS00583 PFKB KINASES 1; PATTERN.
8> PS00584 PFKB KINASES 2; PATTERN.

1> AC PS00001
ID ASN_GLYCOSYLATION; PATTERN.
DE N-glycosylation site.
PA N-{P}-[ST]-{P}

NFSG 107-110
NASG 233-236

```
1       10        20        30        40
msrrvatitlnpaydlvgfcpeiergevnlvkttglhaagkginvakvlk 50
dlgidvtvggflgkdnqdgfqqlfselgianrfqvvqgrtrinvkltekd 100
gevtdfNFSGfevtpadwerfvtdslswlgqfdmvcvsgslpsgvspeaf 150
tdwmtrlrsqcpciifdssrealvaglkaapwlvkpnrreleiwagrklp 200
emkdvieaahalreqgiahvvislgaegalwvNASGewiakppsvdvvst 250
vgagdsmvggliygllmressehtlrlatavaalavsqsnvgitdrpqla 300
ammarvdlqpfn 312
```

Back to the top

7> AC PS00583
ID PFKB_KINASES_1; PATTERN.
DE pfkB family of carbohydrate kinases signature 1.
PA [AG]-G-x(0,1)-[GAP]-x-N-x-[STA]-x(6)-[GS]-x(9)-G

AGKGINVAKVLKDLGIDVTVGGFLG 39-63

```
1       10        20        30        40
msrrvatitlnpaydlvgfcpeiergevnlvkttglhaAGKGINVAKVLK 50
DLGIDVTVGGFLGkdnqdgfqqlfselgianrfqvvqgrtrinvkltekd 100
gevtdfnfsgfevtpadwerfvtdslswlgqfdmvcvsgslpsgvspeaf 150
tdwmtrlrsqcpciifdssrealvaglkaapwlvkpnrreleiwagrklp 200
emkdvieaahalreqgiahvvislgaegalwvnasgewiakppsvdvvst 250
vgagdsmvggliygllmressehtlrlatavaalavsqsnvgitdrpqla 300
ammarvdlqpfn 312
```

Back to the top

8> AC PS00584
ID PFKB_KINASES_2; PATTERN.
DE pfkB family of carbohydrate kinases signature 2.
PA [DNSK]-[PSTV]-x-[SAG](2)-[GD]-D-x(3)-[SAGV]-[AG]-

STVGAGDSMVGGLI 249-262

```
1       10        20        30        40
msrrvatitlnpaydlvgfcpeiergevnlvkttglhaagkginvakvlk 50
dlgidvtvggflgkdnqdgfqqlfselgianrfqvvqgrtrinvkltekd 100
gevtdfnfsgfevtpadwerfvtdslswlgqfdmvcvsgslpsgvspeaf 150
tdwmtrlrsqcpciifdssrealvaglkaapwlvkpnrreleiwagrklp 200
emkdvieaahalreqgiahvvislgaegalwvnasgewiakppsvdvvST 250
VGAGDSMVGGLIygllmressehtlrlatavaalavsqsnvgitdrpqla 300
ammarvdlqpfn 312
```

Pattern Search Result (II)

- **One Query Pattern Against Sequence Database**

Your input pattern
AC PS00888
ID CNMP_BINDING_1; PATTERN.
DE Cyclic nucleotide-binding domain signature 1.
[LIVM]-[VIC]-x(2)-G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x(2)-G
matches **80** sequences in the **PIR-PSD**
Displaying page **1** (up to 50 entries per page) Next page

Sort by NF ID Re-Load Search Protein Name

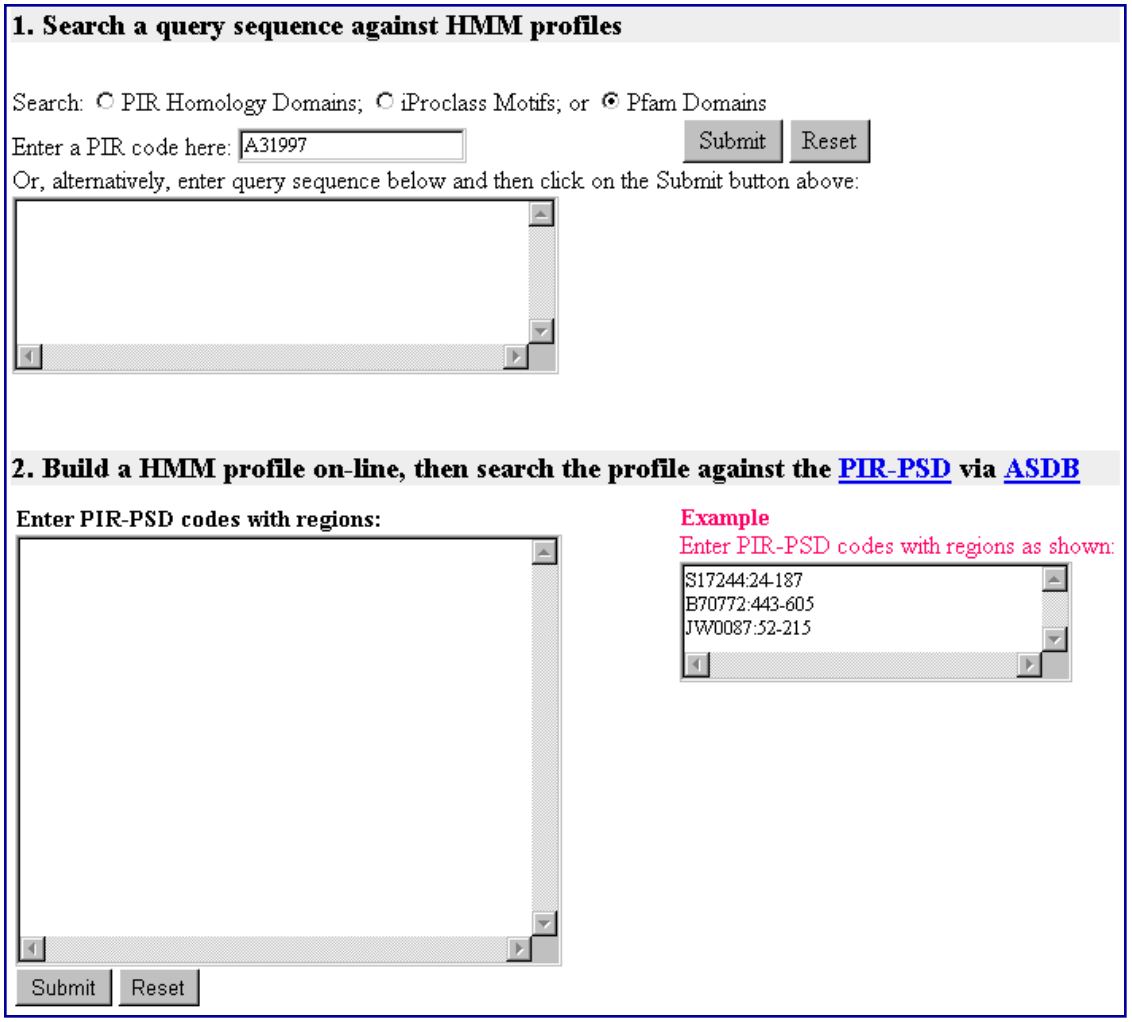
To perform a sequence analysis, make your selection below, check the box before the sequence ID, and then click on the Submit button. submit
BLAST; FASTA; Pattern Match; HMM Search; Multiple alignment; Domain Display;

ID	Protein Name	Organism	Taxon Group	Superfamily	Family	length	Match Range
NREF:NF00013325 iProClass:NF00013325 PIR-PSD:OKFF1R SP/TR:KAPR_DROME	cAMP-dependent protein kinase regulatory chain type I (DRI class I to class IV)	Drosophila melanogaster(fruit fly)	Euk/Animal	SF000548	FAM0000610	377	159-175: IIQQGDEGDNFYVIDVG 277-293: IVKQGAAGDDFYIILEG
NREF:NF00014139 iProClass:NF00014139 PIR-PSD:D34106	protein kinase (EC 2.7.1.37), cGMP-dependent 2, type cD5	Drosophila melanogaster(fruit fly)	Euk/Animal		FAM0009227	614	510-526: IVRQGARGDTFFIISKG
NREF:NF00014387 iProClass:NF00014387 PIR-PSD:T08418	protein kinase (EC 2.7.1.37), cGMP-dependent, type cD5	Drosophila melanogaster(fruit fly)	Euk/Animal		FAM0009227	934	511-527: IVRQGARGDTFFIISKG
NREF:NF00018662 iProClass:NF00018662 PIR-PSD:A34106 SP/TR:KGP1_DROME	cGMP-dependent protein kinase, isozyme 1 (EC 2.7.1.37) (CGK)	Drosophila melanogaster(fruit fly)	Euk/Animal	SF000559	FAM0000617	768	331-347: IIRQGTAGDSFFLISQG
NREF:NF00019403 iProClass:NF00019403 PIR-PSD:B34106 SP/TR:KGP2_DROME	cGMP-dependent protein kinase, isozyme 2 forms cD4/T1/T3A/T3B (EC 2.7.1.37) (CGK) (Foraging protein)	Drosophila melanogaster(fruit fly)	Euk/Animal	SF000559	FAM0000617	1088	547-563: IIKEGDVGSIVYVMEDG 665-681: IVRQGARGDTFFIISKG

Profile Method

- Profile: A Table of Scores to Express Family Consensus Derived from Multiple Sequence Alignments
  - Num of Rows = Num of Aligned Positions
  - Each row contains a score for the alignment with each possible residue.
- Profile Searching
  - Summation of Scores for Each Amino Acid Residue along Query Sequence
  - Higher Match Values at Conserved Positions

PIR HMM Domain/Motif Search



- From Text/Sequence Search Result or HMM Search Interface
- HMMER Model Building & Sequence Search
- Search One Query Protein Against All HMMs
- Search One HMM Against Sequence DB

HMM Search Result (I)

- One Query Protein Against All Pfam HMMs

Query: A31997 IMP dehydrogenase (EC 1.1.1.205) II - human

Scores for Pfam Classification

Model	Description	Score	E-value	N
IMPDH_C	IMP dehydrogenase / GMP reductase C terminus	475.0	6.2e-139	1
IMPDH_N	IMP dehydrogenase / GMP reductase N terminus	201.1	1.7e-56	1
CBS	CBS domain	87.3	3.1e-22	2
NPD	2-nitropropane dioxygenase	-116.3	0.09	1

Parsed for domains

Model	Repeat	Seq-f	Seq-t		hmm-f	hmm-t		Score	E-value
IMPDH_N	1/1	21	110	..	1	88	[]	201.1	1.7e-56
CBS	1/2	112	168	..	1	54	[]	38.8	1.2e-07
CBS	2/2	179	232	..	1	54	[]	48.6	1.4e-10
NPD	1/1	52	463	..	1	408	[]	-116.3	0.09
IMPDH_C	1/1	262	483	..	1	231	[]	475.0	6.2e-139

Alignments of top-scoring domains

```
IMPDH_N: domain 1 of 1, from 21 to 110: score 201.1, E = 1.7e-56
                     *->eklfk..egLTFDDvLLlPaysdvlPkeVdlstrLTknIkLniPlvS
                        ++lf  ++gLT++D+L+lP+y+d++++ Vdl++ LTk+I+L++PlvS
           A319     21    QQLFNcgDGLTYNDFLILPGYIDFTADQVDLTSALTKKITLKTPLVS 67

                        saMDTVTEseMAiaMArlGGiGiIHkNmsieeQAeeVrkVKky<-*
                        s+MDTVTE+ MAiaMA+ GGiG+IH+N+++e QA+eVrkVKky
           A319     68 SPMDTVTEAGMAIAMALTGGIGFIHHNCTPEFQANEVRKVKKY    110

CBS: domain 1 of 2, from 112 to 168: score 38.8, E = 1.2e-07
                     *->mimqedvvtvspttttleealnllrehgisrlpVvded...grlvGiv
                        +++++d+v++sp + ++++++   +hg+ ++p++d  + + rlvGi+
           A319    112    QGFITDPVVLSPKDRVRDVFEAKARHGFCGIPITDTGrmgSRLVGII 158
```

HMM Search Result (II)

- Search User-Built HMM Against Protein Sequence DB
- Input Sequences (Optional Residue Ranges) -> Multiple Sequence Alignment -> Model Building -> HMM Search

Please verify the alignment below and then click on here to start the HMM procedures.
You could modify the alignment. But do not change the format and PIR codes.

```
CLUSTAL W (1.8) multiple sequence alignment


S17244_24-187       GCTIWLTGLSASGKSTIACALEQLLQKNLSAYRLDGDNIRFGLNKDLGFSEKDRNENIR
B70772_443-605      GKTVWFTGLSGSGKSSVAMLVERKLLEKGISAYVLDGDNLRHGLNADLGFSMADRAENLR
JW0087_52-215       GCTVWLTGLSGAGKTTVSMALEEYLVCHGIPCYTLDGDNIRQGLNKNLGFSPEDREENVR
                    :*. *: :.:..* *****:* *** :****  ** **:*

                    ISPYRVDRDRARELHKEAGLKFIEIFVDVPLEVAEQRDPK
                    ISPLAEHRALARKVHADAGIDFFEVFCDTPLQDCERRDPK
                    ISPYTQDRNNARQIHEGASLPFFEVFVDAPLHVCEQRDVK
                    ***   .*  **::*  *.: *:*:* *.**. .*:** *

                    YEAPKAPELHLRTDQKTVEECATI
                    YQRPKNPDLRLTPD-RSIDEQAQE
                    YEKPEAPELVLKTDSCDVNDCVQQ
                    *: *: *:* * .*   ::: .
```

Click here to show the HMM model. Click on the Score for the alignment and the Title for the similar sequances

Scores for complete sequences (score includes all domains):

ID	Super Fam	Family	Title	Score	E-value	N
JW0087	SF001612	FAM020020	adenylyl-sulfate kinase (EC 2.7.1.25)	501.9	2.5e-149	1
S17244	SF000544	FAM005826	adenylyl-sulfate kinase (EC 2.7.1.25)	491.3	3.9e-146	1
B70772	SF003009	FAM003167	probable adenylyl-sulfate kinase (EC 2.7.1.25)	483.7	8e-144	1
JC4383	SF001612	FAM020020	adenylyl-sulfate kinase (EC 2.7.1.25)	408.5	3.5e-121	1
T50101	SF000544	FAM005826	adenylylsulfate kinase [imported]	381.1	6e-113	1
T24918	SF001612	FAM020020	3'-phosphoadenosine-5'-phosphosulfate synthetase	334.4	7e-99	1
A83836	SF000544	FAM005826	adenylylsulfate kinase BH1489 [imported]	315.3	3.8e-93	1
E96912	SF000544	FAM005826	adenylylsulfate kinase [imported]	311.3	6.1e-92	1
A69839	SF000544	FAM005826	adenylylsulfate kinase homolog yisZ	305.3	3.9e-90	1
A87433	SF003009	FAM003167	hypothetical protein CC1482 [imported]	293.5	1.4e-86	1