

# Lecture 2

Beyond database searching: How do we turn  
our results into knowledge?

Patricia C. Babbitt  
Associate Professor  
Department of Biopharmaceutical Sciences  
& Biological and Medical Information Sciences Graduate Group

August 22, 2002

- Some Basic Principles of Molecular Evolution
- Evaluation using Multiple Alignments
- Finding and Analyzing Motifs
- New Directions in Bioinformatics

# Molecular Evolution

Highly relevant but we only have time to  
mention some  
very basic issues

# References

Saier, M.H. Jr. "Phylogenetic approaches to the identification and characterization of protein families and superfamilies"

Labedan, B. & Riley, M. "Gene products of E.coli: Sequence comparisons and common ancestries"

Green, P. et al. "Ancient conserved regions in new gene sequences and the protein databases"

Murzin, A.G. "How far divergent evolution goes in proteins"

Textbooks:

Fundamentals of Molecular Evolution, Li & Graur, Sinauer Associates, 2nd Ed. (1999)

Molecular Systematics, D.M. Hillis & C. Moritz, Eds., Sinauer Associates (1990)

# Web Resources

- **Useful Lists**

<http://www.mcb.harvard.edu/BioLinks/Evolution.html>  
<http://darwin.eeb.uconn.edu/molecular-evolution.html>

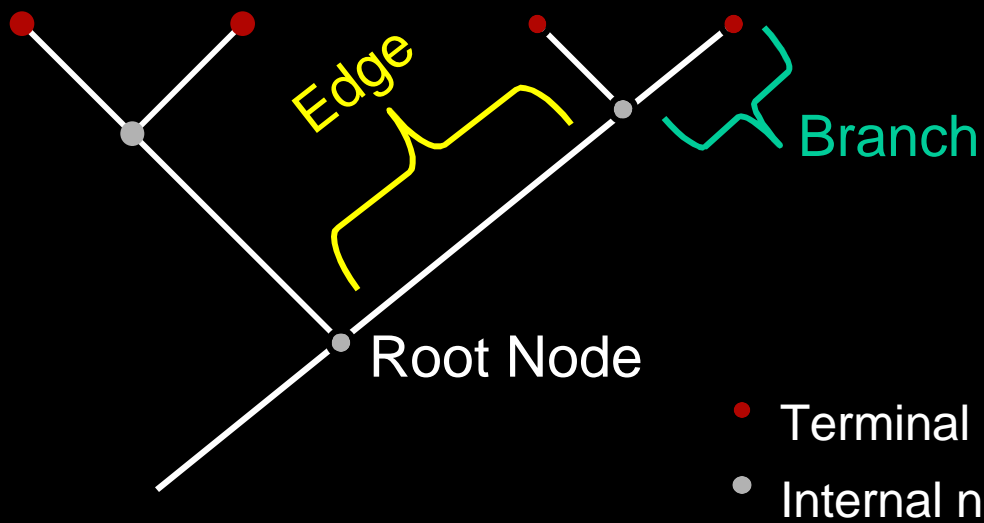
- **Tree of Life site**

<http://phylogeny.arizona.edu/tree/phylogeny.html>

- **A protocol to get you started**

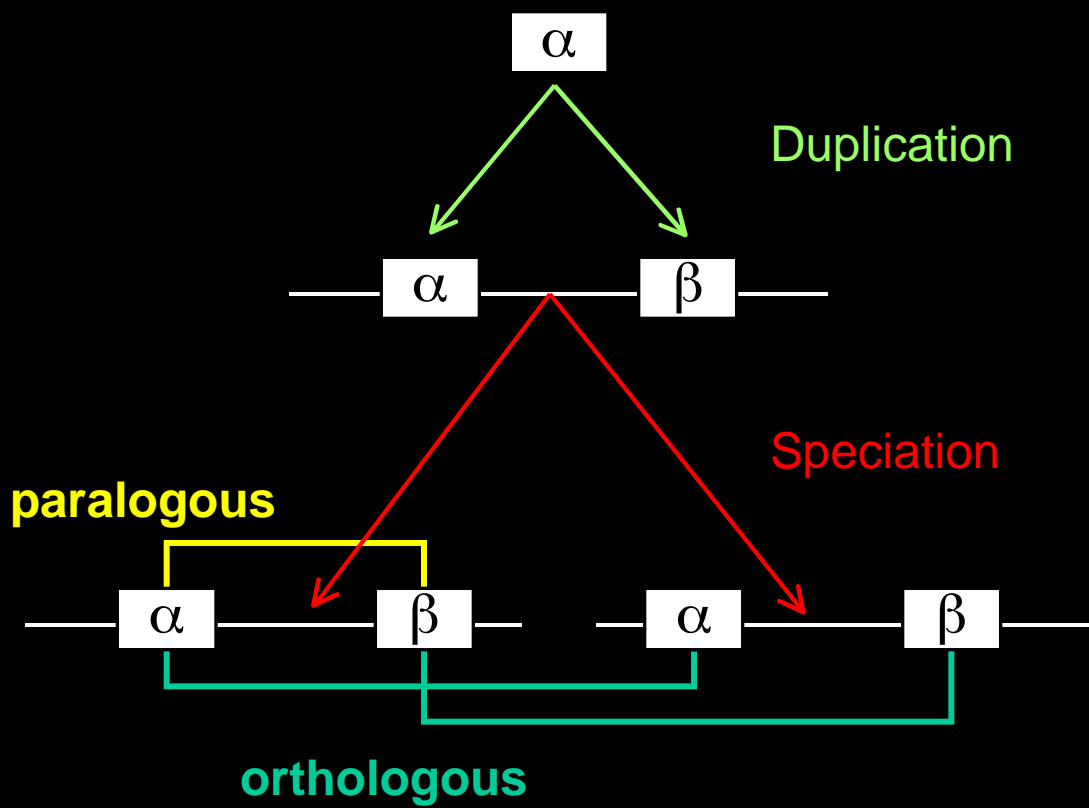
<http://www.infobiogen.fr/docs/MAcours/phylogeny.html>

# Tree (Network) Nomenclature



## Definitions

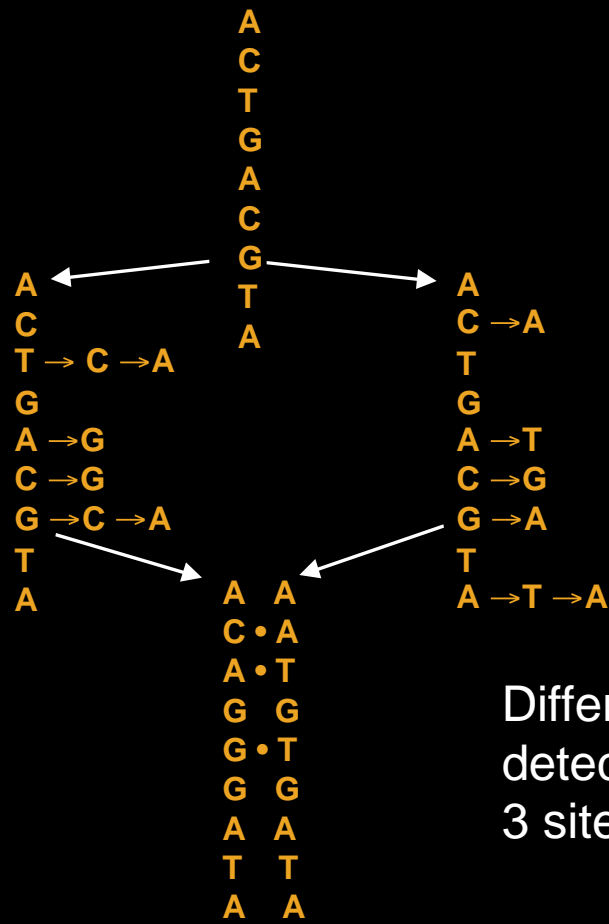
- Homology: Sharing a common ancestor, may have similar or dissimilar functions
- Analogy: Performing a common function but no common ancestry
- Convergence: Performing the same function, having similar structural characteristics, but do not share a common ancestor
- Paralogy: Sequence similarity between the descendants of a duplicated ancestral gene
- Orthology: Sequence similarity as a consequence of a speciation event



## Important principles

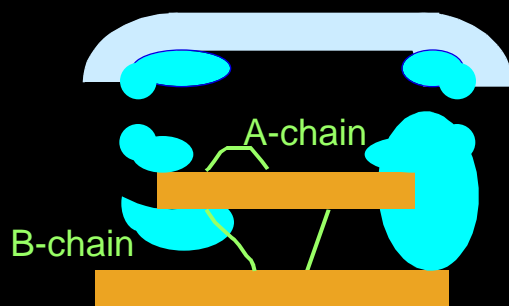
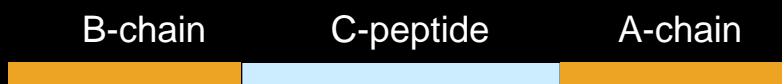
- Evolutionary history is accessed only through contemporary species and molecules
- The basic models for substitution are generally robust for sequences 80% identical (nucleotide level), e.g., not highly diverged
- General assumptions of the models
  - Changes in different copies of genes are independent
  - Changes at each site are independent
  - All sites change at the same rate
  - All bases occur at equal frequencies (corrected in later models to come a little closer to reality)

12 mutations  
accumulated



Differences  
detected at only  
3 sites

- Different domains within a single protein evolve at different rates



Mature insulin

$$r = 0.97 \times 10^{-9}/\text{site}/\text{year}$$

$$r = 0.13 \times 10^{-9}/\text{site}/\text{year}$$

# Evaluation using Multiple Alignments

# References on multiple alignment tools

McClure, "Comparative analysis of multiple protein sequence analysis methods"

Thompson et al., "ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice"

(MSA) Lipman et al., "A tool for multiple sequence alignment"

Notredame & Higgins, "SAGA: Sequence alignment by genetic algorithm"

(PIMA) Smith & Smith "Automatic generation of primary sequence patterns from sets of related protein sequences"

See also:

(MACAW) Schuler, G.D., Altschul, S.F., Lipman, D.J. (1991) "A workbench for multiple alignment construction and analysis," *Proteins* 9, 180-90

(PILEUP) Feng, D.F. & Doolittle, R.F. "Progressive sequence alignment as a prerequisite to correct phylogenetic trees" (1987) *J. Mol. Evol.* 25, 351-60

# Evaluation of sequence relationships using multiple alignments

- Screening for membership in a family/superfamily
- Identification of conserved elements important to function
- Distinguishing global vs. local patterns of similarity characteristic of the structural scaffold
- Determination of the level and sites of variability across the members of subgroups/families/superfamilies

- Multiple alignments are more informative than pairwise comparisons

BLASTP 1.4.9

Query= TITLE: URF  
(269 letters)

Database: Non-redundant SwissProt sequences  
49,825 sequences; 17,390,645 total letters.

		Smallest Sum Probabil.
sp P24162	ENOYL-COA HYDRATASE HOMOLOG (ORF257)...	6.1e-31
sp P34559	PROBABLE ENOYL-COA HYDRATASE, MITOCH...	5.2e-29
sp P14604	ENOYL-COA HYDRATASE, MITOCHONDRIAL P...	3.1e-28
sp P30084	ENOYL-COA HYDRATASE, MITOCHONDRIAL P...	1.3e-24
sp P23966	NAPHTHOATE SYNTHASE (DIHYDROXYNAPHTH...	2.3e-21

•  
•  
•

BLASTP 1.4.9

Query= TITLE: Urf  
(269 letters)

Database: Non-redundant SwissProt sequences  
49,825 sequences; 17,390,645 total letters.

	Smallest Sum Probabil.
sp P24162 ENOYL-COA HYDRATASE HOMOLOG (ORF257)...	6.1e-31
sp P34559 PROBABLE ENOYL-COA HYDRATASE, MITOCH...	5.2e-29
sp P14604 ENOYL-COA HYDRATASE, MITOCHONDRIAL P...	3.1e-28
sp P30084 ENOYL-COA HYDRATASE, MITOCHONDRIAL P...	1.3e-24
sp P23966 NAPHTHOATE SYNTHASE (DIHYDROXYNAPHTH...	2.3e-21

•  
•  
•

```
>sp|P24162|ECHH_RHOCA ENOYL-COA HYDRATASE HOMOLOG (ORF257). >pir||S19026
    enoyl-CoA hydratase homolog - Rhodobacter capsulatus >gi|45984
    (X60194) enoyl-CoA hydratase homologue [Rhodobacter capsulatus]
    Length = 257
```

```
Score = 207 (96.1 bits), Expect = 6.1e-31, Sum P(3) = 6.1e-31
Identities = 51/137 (37%), Positives = 71/137 (51%)
```

```
Query:   89 WHQMIHKIIRVKRPVLAANGVAAGGGLGISLASDMAICADSAKFVCAWHTIGIGNDTAT 148
      +  ++ I      PVLAA+NG AAG G  ++LA+D+ I A SA F+ A+  IG+  D
```

```
Sbjct:   83 YEPLLQAIYSCPLPVLAAVNGAAAGAGANLALAADVIAAQSAAFMQAFTRIGLMPDAGG 142
```

```
Query:   149 SYSLARIVGMRRAMEMLTNRITYPEEAKDWGLVSRVYPKDEFREVAWKVARELAAAPTH 208
      ++ L R VGM RAM + L    + EEA  GL+   P  +F      A LA P+
```

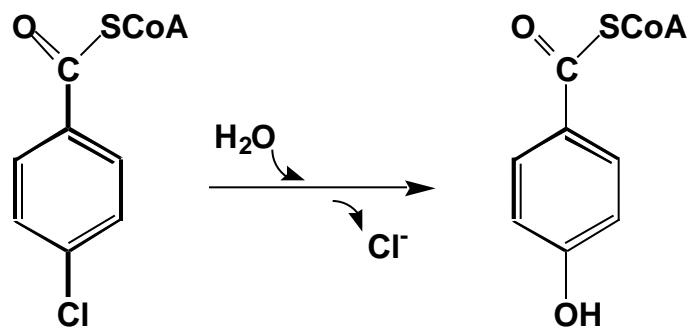
```
Sbjct:   143 TWWLPRQVGMARAMGMALFAEKIGAEAAARMGLIWEAVPDVDFEHHWRARAAHLARGPSA 202
```

```
Query:   209 LNVMAKERFHAGWMNPV 225
```

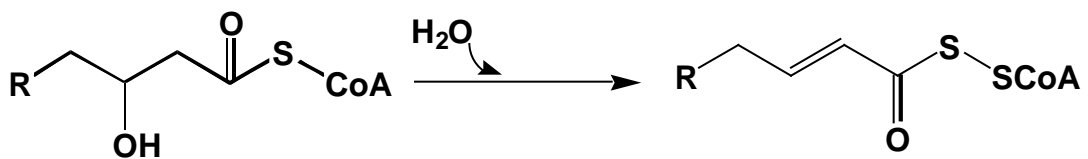
```
      K+ FHAG  NP+
```

```
Sbjct:   203 AFAAVKKAFFHAGLSNPL 219
```

### 4-Chlorobenzoate Dehalogenase



### Enoyl-CoA Hydratase



- A multiple alignment distinguishes the dehalogenase from the enoyl Co-A hydratase family

Dehalogenases	GGGLGISLASDMAICADSAKFVCAWHTIGIGNDTAT
	GGGLGMSLACDLAVCTDRATFLPAWMSIGIANDASS
Enoyl CoA Hydratases	GGGCELAMMCDIIYAGEKAQFGQPEILLGTIPGAGG
	GGGNELAMMCDIIYAGEKARFGQPEINIGTIPGAGG
	GGGCELAMMCDFIIASETAKFGLPEITLGVIPGMGG
	GAGCELALLCDVVVAGENARFGLPEITLGIIMPAGG

## Multiple alignments provide more information than pairwise alignments

- Useful to confirm distant relationships
- Provides a context for interpreting patterns of similarity and difference
- "Speciation" over alignment space helps to connect and confirm widely degenerate motifs

Query= /phosphonatase/phosBc.gcg (302 letters)

Database: swissprot  
77,273 sequences; 27,815,109 total letters.

Sequences producing High-scoring Segment Pairs:				High	Smallest	
				Score	Sum	Probability
					P(N)	N
sp	P77247	YNIC_ECOLI	HYPOTHETICAL 24.3 KD PROTEIN IN PFKB...	116	2.2e-05	1
sp	O67359	GPH_AQUAE	PHOSPHOGLYCOLATE PHOSPHATASE (PGP)	106	0.00030	1
sp	O06995	PGMB_BACSU	PUTATIVE BETA-PHOSPHOGLUCOMUTASE (BE...	97	0.0039	1
sp	P31467	YIEH_ECOLI	HYPOTHETICAL 24.7 KD PROTEIN IN TNAB...	94	0.0082	1
sp	P44755	GPH_HAEIN	PHOSPHOGLYCOLATE PHOSPHATASE (PGP)	93	0.011	1
sp	P54607	YHCW_BACSU	HYPOTHETICAL 24.7 KD PROTEIN IN CSPB...	89	0.030	1
sp	P32662	GPH_ECOLI	PHOSPHOGLYCOLATE PHOSPHATASE (PGP)	87	0.067	1

~ 21% identical

PGPhos      - - - - - M P G **V** **V** **F** **D** **L** **D** **G** **T** **L** **V** H S A P D I H A A **V** N K  
Phosphon    M D R M K I E A **V** **I** **F** **D** W A **G** **T** **T** **V** D Y G C F A P L E **V** F M

PGPhos      A L A E E G **G** A P F T L A **E** I T G F I **G** - - N G V P **V** L I Q  
Phosphon    E I F H K R **G** V A I T A E **E** A R K P M **G** L L K I D H **V** R V T

PGPhos      R V L A A R G **E** A P D A H **R** **Q** A E L Q G R F M A H **Y** **E** A D P  
Phosphon    E M P R I A S **E** W N R V F **R** **Q** L P T E A D I Q E M **Y** **E** E F E

PGPhos      A T **L** T S V Y **P** - - - - - - **G** A E A A **I** R H **L** R A E **G** W R  
Phosphon    E I **L** F A I L **P** R Y A S P I N **G** V K E V **I** A S **L** R E R **G** I K

PGPhos      **I** **G** L C **T** N K P V **G** A S **R** Q I L S L F - - - G L **L** E L F - -  
Phosphon    **I** **G** S T **T** - - - - **G** Y T **R** E M M D I V A K E A A **L** Q G Y K P

PGPhos      **D** A I I G G E S L **P** Q R K **P** D P A P L R A T A A A L N - - -  
Phosphon    **D** F L V T P D D V **P** A G R **P** Y P W M S Y K N A M E L **L** G V Y P

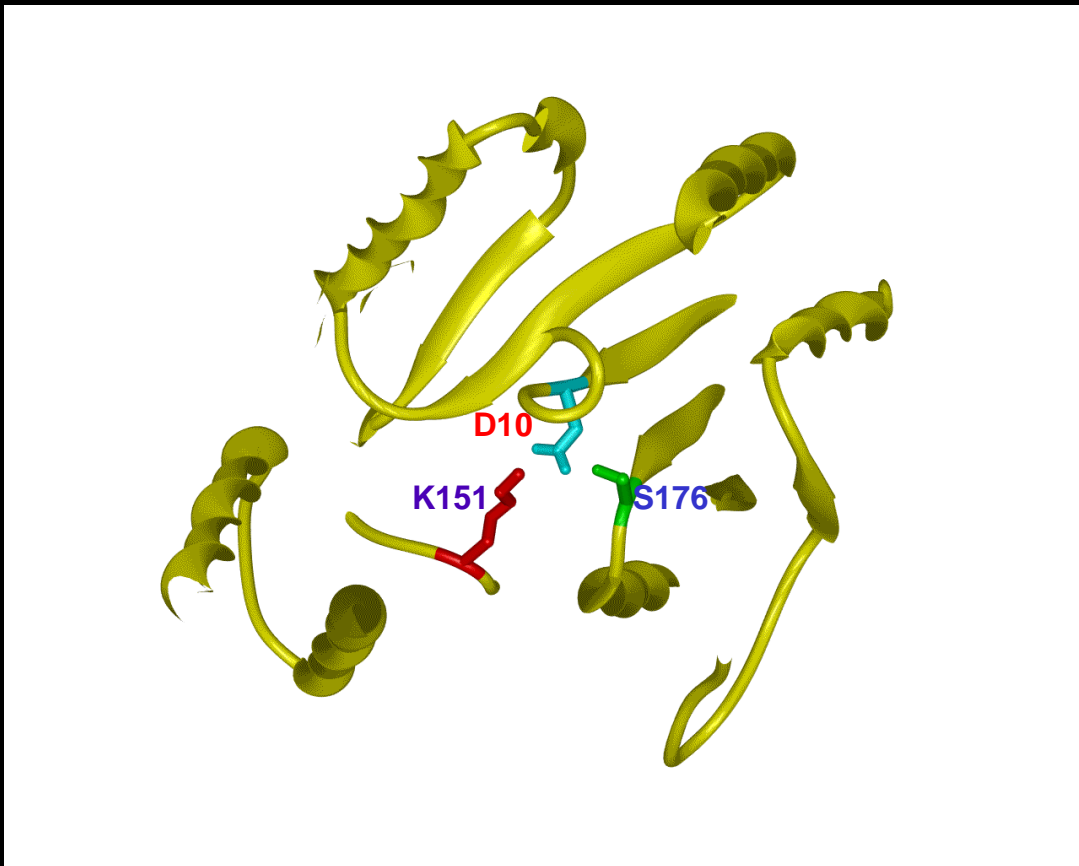
PGPhos      E E V V L Y **V** **G** **D** S E V **D** A A T A E A **A** **G** L R F A L F T E **G**  
Phosphon    M N H M I K **V** **G** **D** T V S **D** M K E G R N **A** **G** M W T V G V I L **G**

PGPhos      Y R H A P V - - **H** **E** L P H H G L F S H H D E L Q D L L R **R** L  
Phosphon    S S E L G L T E E **E** V E N M D S V E L R E K I E V V R N **R** F

	<b>10</b> *		<b>151</b> *		<b>176</b> *		
Cu++ATPase. Ec	LDTVV	<b>F D</b> KTGT	LTEG	VIAGVLPDG	<b>K</b> AEAIKHL	AMVG	<b>D</b> GINDAPAL
Cu++ATPase. Hs	VKVVV	<b>F D</b> KTGT	ITHG	VFAEVLPSH	<b>K</b> VAKVKQL	AMVG	<b>D</b> GINDSPAL
Ca++ATPase. At	ATTIC	<b>S D</b> KTGT	LTTN	VMARSSPMD	<b>K</b> HTLVRL	AVTG	<b>D</b> GTNDAPAL
Urf. Mj	KVAIV	<b>F D</b> SAGT	LVKI	E - - AHQEL	<b>K</b> RDLIRNL	IMVG	<b>D</b> GANDVPAM
PhosSerPhos. Hs	ADAVC	<b>F D</b> VDST	VI RE	TAE - SGGKG	<b>K</b> VI KLLKE	IMI G	<b>D</b> GATDMEAC
2-D0-6-PPhos. Sc	VDLCL	<b>F D</b> LDGT	IVST	ITGFDVKNG	<b>K</b> PDPEGYS	VVFE	<b>D</b> APVGIKAG
DL-Gly-3-Phos. Sc	INAAL	<b>F D</b> VDGT	IIIS	ITANDVKQG	<b>K</b> PHPEPYL	VVFE	<b>D</b> APAGIAAG
Phosphon. Pa	LQAAI	<b>L D</b> WAGT	VVDF	ATDEV - PNG	<b>R</b> PWPAAAL	VKVD	<b>D</b> TWPGILEG
Phosphon. St	IHAVI	<b>L D</b> WAGT	TVDF	ATDDLAAGG	<b>R</b> PGPWMAAL	VKVD	<b>D</b> AAPGISEG
Phosphon. Bc	IEAVI	<b>F D</b> WAGT	TVDY	TPDDV - PAG	<b>R</b> PYPWMSY	IKVG	<b>D</b> TVSDMKEG
PhosGlycolPhos. Rs	MPGVV	<b>F D</b> LDGTLVHS		IGGESLPQR	<b>K</b> PDPAPLA	LYVG	<b>D</b> SEVDAATA
NtermDom. IGPD. Pp	VQALL	<b>L D</b> MDGV	MAEV	LED CPP - - -	<b>K</b> PSPEPIL	AMVG	<b>D</b> TVDDIIAG
B-PhosGlucoMut. Ll	FKAVL	<b>F D</b> LDGV	ITDT	AEVAAS - - -	<b>K</b> PAPPDIFI	IGLE	<b>D</b> SQAGIQAI
Hal oAci dDehal . PspYL	IKGIA	<b>F D</b> LYGT	LFDV	LSVDPVQVY	<b>K</b> PDNRVYE	LFVS	<b>S</b> NAWDATGA
NtermDomEpoxHyd. Hs	LRAAV	<b>F D</b> LDGV	LALP	I ESCQVGMV	<b>K</b> PEPQIYK	VFLD	<b>D</b> IGANLKPA
EnolasePhos. Ko	IRAI V	<b>T D</b> IEGT	TS DI	FD - - TLVGA	<b>K</b> REAQSYR	LFLS	<b>D</b> IHQELDAA

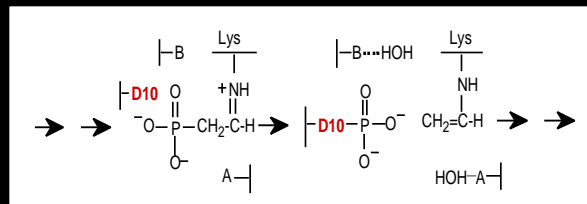
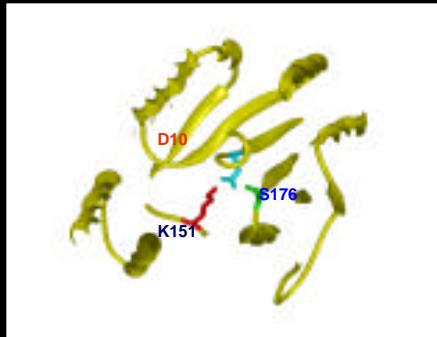
	10 *		151 *		176 *			
Cu++ATPase. Ec	LDTVV	F D KTGT	LTEG	VIAGVLPDG	K AEAI KHL	AMVG	D	GINDAPAL
Cu++ATPase. Hs	VKVVV	F D KTGT	ITHG	VFAEVLPSH	K VAKVKQL	AMVG	D	GINDSPAL
Ca++ATPase. At	ATTIC	S D KTGT	LTTN	VMARSSPMD	K HTLVRL	AVTG	D	GTNDAPAL
Urf. Mj	KVAIV	F D SAGT	LVKI	E - - AHQEL	K RDLIRNL	IMVG	D	GANDVPAM
PhosSerPhos. Hs	ADAVC	F D VDST	VI RE	TAE - SGGKG	K VI KLLKE	IMI G	D	GATDMEAC
2-D0-6-PPhos. Sc	VDLCL	F D LDGT	IVST	ITGFDVKNG	K PDPEGYS	VVFE	D	APVGIKAG
DL-Gly-3-Phos. Sc	INAAL	F D VDGT	IIIS	ITANDVKQG	K PHPEPYL	VVFE	D	APAGIAAG
Phosphon. Pa	LQAAL	L D WAGT	VVDF	ATDEV - PNG	R PWPAQAL	VKVD	D	TWPGILEG
Phosphon. St	IHAVI	L D WAGT	TVDF	ATDDLAAGG	R PGPWMAL	VKVD	D	AAPGISEG
Phosphon. Bc	IEAVI	F D WAGT	TVDY	TPDDV - PAG	R PYPWMSY	IKVG	D	TVSDMKEG
PhosGlycol Phos. Rs	MPGVV	F D LDGTLVHS		IGGESLPQR	K PDPAPLA	LYVG	D	SEVDAATA
NtermDom. IGPD. Pp	VQALL	L D MDGV	MAEV	LED CPP - -	K PSPEPIL	AMVG	D	TVDDIIAG
B-PhosGlucMut. Ll	FKAVL	F D LDGV	ITDT	A EVAAS - -	K PAPPDIFI	IGLE	D	SQAGIQAI
Hal oAci dDehal. PspYL	IKGIA	F D LYGT	LFDV	LSVDPVQVY	K PDNRVYE	LFVS	S	NAWDATGA
NtermDomEpoxyHyd. Hs	LRAAV	F D LDGV	LALP	I ESCQVGMV	K PEPQIYK	VFLD	D	IGANLKPA
EnolasePhos. Ko	IRAIV	T D IEGT	TSDI	FD - - TLVGA	K REAQSYR	LFLS	D	IHQELDAA

## Active site of haloacid dehalogenase



	<b>10</b> *	<b>151</b> *	<b>176</b> *
Cu++ATPase. Ec	LDTVV <b>F D</b> KTGTLTEG	VIAGVLPDG <b>K</b> A EAIKHL	AMVG <b>D</b> GINDAPAL
Cu++ATPase. Hs	VKVVV <b>F D</b> KTGTITHG	VFAEVLPSH <b>K</b> VAKVKQL	AMVG <b>D</b> GINDSPAL
Ca++ATPase. At	ATTIC <b>S D</b> KTGTLTTN	VMARSSPMD <b>K</b> HTLVRL	AVTG <b>D</b> GTNDAPAL
Urf. Mj	KVAIV <b>F D</b> SAGTLVKI	E - - AHQEL <b>K</b> RDLIRNL	IMVG <b>D</b> GANDVPAM
<b>PhosSerPhos. Hs</b>	<b>A D A V C F D V D S T V I R E</b>	<b>T A E - S G G K G K V I K L L K E</b>	<b>I M I G D G A T D M E A C</b>
2-D0-6-PPhos. Sc	VDLCL <b>F D</b> LDGTIVST	ITGFDVKNG <b>K</b> PDPEGYS	VVFE <b>D</b> APVGIKAG
DL-Gly-3-Phos. Sc	INAAL <b>F D</b> VDGTIIIS	ITANDVKQG <b>K</b> PHPEPYL	VVFE <b>D</b> APAGIAAG
Phosphon. Pa	LQAAL <b>L D</b> WAGTVVDF	ATDEV-PNG <b>R</b> PWPAQAL	VKVD <b>D</b> TWPGILEG
Phosphon. St	IHAVI <b>L D</b> WAGTVVDF	ATDDLAAG <b>R</b> PGPWML	VKVD <b>D</b> AAPGISEG
Phosphon. Bc	IEAVI <b>F D</b> WAGTVVDY	TPDDV-PAG <b>R</b> PYPWMSY	IKVG <b>D</b> TVSDMKEG
PhosGlycolPhos. Rs	MPGVV <b>F D</b> LDGTLVHS	IGGESLPQR <b>K</b> PDPAPLA	LYVG <b>D</b> SEVDAATA
NtermDom. IGPD. Pp	VQALL <b>L D</b> MDGVMAEV	LED CPP - - <b>K</b> PSPEPIL	AMVG <b>D</b> TVDDIIAG
B-PhosGlucoMut. Ll	FKAVL <b>F D</b> LDGVIITDT	AEVAAS - - <b>K</b> PAPDIFI	IGLE <b>D</b> SQAGIQAI
Hal oAci dDehal. PspYL	IKGIA <b>F D</b> LYGTLFDV	LSVDPVQVY <b>K</b> PDNRVYE	LFVS <b>S</b> NAWDATGA
NtermDomEpoxyHyd. Hs	LRAAV <b>F D</b> LDGVLALP	IESCQVGMV <b>K</b> PEPQIYK	VFLD <b>D</b> IGANLKPA
<b>EnolasePhos. Ko</b>	<b>I R A I V T D I E G T T S D I</b>	<b>F D - - T L V G A K R E A Q S Y R</b>	<b>L F L S D I H Q E L D A A</b>

	10	151	176
Cu++ATPase. Ec	LDTVV <b>F</b> DKTGTLTEG	VIAGVLPDG <b>K</b> A <b>E</b> AIKHL	AMVG <b>D</b> GINDAPAL
Cu++ATPase. Hs	VKVVV <b>F</b> DKTGTLTHG	VFAEVLPSH <b>K</b> V <b>A</b> KVKQL	AMVG <b>D</b> GINDSPAL
Ca++ATPase. At	ATTIC <b>S</b> DKTGTLTTN	VMARSSPMD <b>K</b> H <b>T</b> LVRL	AVTG <b>D</b> GTNDAPAL
Urf. Mj	KVAIV <b>F</b> DSAGTLVKI	E--AQEL <b>K</b> R <b>D</b> LIRNL	IMVG <b>D</b> GANDVPAM
PhosSerPhos. Hs	ADAVC <b>F</b> DVDSIVIRE	TAE--SGG <b>K</b> G <b>K</b> V <b>I</b> KLLKE	IMIG <b>D</b> GATDMEAC
2-D0-6-PPhos. Sc	VDLCL <b>F</b> LDGTLIVST	ITGFVKN <b>K</b> G <b>P</b> DPEGYS	VVFE <b>D</b> APVGIKAG
DL-Gly-3-Phos. Sc	INAA <b>L</b> <b>F</b> DVDGTLIIS	ITANDVK <b>K</b> G <b>K</b> P <b>H</b> PEPYL	VVFE <b>D</b> APAGIAAG
Phospon. Pa	LQAA <b>L</b> <b>L</b> DWAGTVVDF	ATDEV--P <b>N</b> G <b>R</b> P <b>W</b> PAQAL	VKVD <b>D</b> TWPGILEG
Phospon. St	IHA <b>V</b> I <b>L</b> DWAGTTVDF	ATDDLAA <b>G</b> <b>R</b> P <b>G</b> PWMAL	VKVD <b>D</b> AAPGISEG
Phospon. Bc	IEA <b>V</b> I <b>F</b> DWAGTTVDY	TPDDV--P <b>A</b> G <b>R</b> P <b>Y</b> PWMSY	IKVG <b>D</b> TVSDMREG
PhosGlycol Phos. Rs	MPGV <b>V</b> <b>F</b> LDGTLVHS	IGGESLP <b>Q</b> R <b>K</b> P <b>D</b> PAPLA	LYVG <b>D</b> SEVDAATA
NtermDom. 1GPD. Pp	VQAL <b>L</b> <b>L</b> DMGVMMAEV	LEDCCP-- <b>K</b> P <b>S</b> PEPIL	AMVG <b>D</b> TVDDIIAG
B-PhosGlucoMut. Ll	FKAV <b>L</b> <b>L</b> LDGVIITDT	AEVAAS-- <b>K</b> P <b>A</b> PDI FI	IGLE <b>D</b> SQAGIQAI
Hal oAci dDehal. PspYL	IKGI <b>A</b> <b>F</b> DLYGTLFDV	LSVDPV <b>Q</b> V <b>K</b> P <b>D</b> NRVYE	LFVS <b>S</b> NAWDATGA
NtermDomEpoxyhyd. Hs	LRA <b>A</b> <b>V</b> <b>F</b> LDGVLALP	IESC <b>Q</b> V <b>G</b> M <b>V</b> <b>K</b> P <b>E</b> PQIYK	VFL <b>D</b> I GANLKPA
EnolasePhos. Ko	IRAI <b>V</b> <b>T</b> D <b>I</b> EGTTS DI	FD--TLV <b>G</b> A <b>K</b> R <b>E</b> AQSYR	LFL <b>S</b> <b>D</b> IHQELDAA



## Issues in using multiple alignment information

- What question are you asking when you create a multiple alignment?
  - Example: GPCRs

Close relationships: Muscarinic receptors

Intermediate relationships: Prostaglandin receptors

Distant relationships: Fungal pheromone receptors

# Muscarinic Receptor Sub-types (45-60% identical)

m2	1	-----MNNST-----NSSNNSLALTS P-
m4	1	-----MANFTPVNGSSGNQSVRLVTSS
m1	1	-----MNTSAPPAVSPNITVL--A
m3	1	MTLHNNSTTSP LFPNISSSWIHSPSDAGLPPGTVTHFGSYNVSRAAGNFSSPDGTTD--D
m5	1	-----MEGDSYHNATTVNGTFVNHQ
m2	18	---YKTFVVFIVLVAGSLSLVTIIGNIDVMSIKVNRHLQTVNNYFLFSLACADLLIIG
m4	23	SHNRYEVEKVFIAVFGSLSLVTITVGNIDVMSIKVNRHLQTVNNYFLFSLACADLLIIG
m1	18	P--GKGTPVQVFIAGITFGSLSLATITVGNIDVMSIKVNTLKTVNNYFLFSLACADLLIIG
m3	59	PLGGHTVQVFIAGITFGSLSLATITVGNIDVMSIKVNTLKTVNNYFLFSLACADLLIIG
m5	21	PLERHRLWFVITIAAVDAVVSGLITIVSNVLMISPKVNSQLKTVNNYFLFSLACADLLIIG
m2	74	VFSMNLVYTLTVIGYWPDLGPVVCDDLWLLALDYVVISNASVMNLLIISFDRYFCVTKPLTYPV
m4	83	AFSMNLVYTVIIKGYWPLGAVVCDDLWLLALDYVVISNASVMNLLIISFDRYFCVTKPLTYPA
m1	76	TFSMNLVYTVLLMCHWALGLTACDDLWLLALDYVAISNASVMNLLIISFDRYFSVTRPLSYRA
m3	119	VFSMNLVYTVIIMNRWALGNLACDDLWLLALDYVAISNASVMNLLIISFDRYFSITRPLTYRA
m5	81	IFSMNLVYTVIILMGRWALGLSACDDLWLLALDYVAISNASVMNLLIISFDRYFSITRPLTYRA
m2	134	KRTTKMAGMMIAAAWVLSFI LWAPAILFWQFIVGVRTVE DGE CYIQFESNAAVTFGTAAIA
m4	143	KRTTKMAGMMIAAAWVLSFV LWAPAILFWQFIVGVKRTVDPDNHCYIQFLSNPAVTFGTAAIA
m1	136	KRTTPRRAALMIGLAWLVSEV LWAPAILFWQYLVGERTVLAGQCQYIQFLSQPIITFGTAAIA
m3	179	KRTTKRAGMMIGLAWVISEV LWAPAILFWQYLVGKRTVPPGECFYIQFLSEPTITFGTAAIA
m5	141	KRTPKRAGMMIGLAWLISEV LWAPAILFWQYLVGKRTVPLDECOYIQFLSEPTITFGTAAIA
m2	194	AFYLPVIMTVLYWHISRASKSR IKKDKKEPVANQDPVSPSLVQG-----R
m4	203	AFYLPVIMTVLYWHISLASRSR VHKKHREP GPKEKAKT LAFLKS-----P
m1	196	AFYLPVIMTVLYWHIYRETENRRARELAALQGSSET-----PGKGGGSSSSSERSQP
m3	239	AFYMPVIMTVLYWHIYKETEKRRTKELAGLQASGT EAE ENFVHP TGSSRS CSSYELQQQ
m5	201	AFYLPVIMTVLYWHIYRETEKRRTKDLADLQGS DSVTKAEK-RKP AHRALFRSCLRCP RP
m2	240	IVKPNNNNMFSSDDGLEHNIKQNGK--APRDPVTENCVQGEKEKSSNDSTSVSAV---
m4	249	LMKQSVKPPFGEAR E--LRNGKLEAPFPALP PPRPVADKDSNESSSG SATQN--
m1	247	GAEQSPETPPGR C-- --CRCCRAPRLQAYSWKE--EEDDEGSSLSLSEDEEPG
m3	299	SMKRSNRRKYGRCHF WFTTKSWKPSSEQMDQDHSSSDSNNNDAAAGLENSASSEDEEDIG
m5	260	TLAQRERNQASWSSSR RST--S TTGKPSQATGPSANWAKAEQLTTCSSYPSSEDEEDK-
m2	293	-----ASNMRDDRITQDENTVSTSLGHSKDENS KQT CI
m4	305	-----TKERPATELSTTBATTPAMPAPPLQPRALNPA RWSKI
m1	298	SE-----VVIKMP-----MVDPEAQAPT KQPPR--SS
m3	359	SETRAIYSIVLKLPGHSTILNSTKLPSSDNLQVPE--EELGMVDLERKADKLOAQK--SV
m5	315	PATDPVLQVVKYK SQG-----KESPGEEFSAE--TEETFVKAETEKSDYDTPNYLLS
m2	326	RIGTKTPKSDSCTPTNTTVEVVGSSSQNGDEKQNI VARKIVKMT-----K
m4	343	QIVTKQTGNECVTA--IEIVPATPAGMRPAAN-VARKFASIA-----R
m1	323	PNTVKKR P-----TKKGRDRAGK GQK-P-----RGKE
m3	415	DDGGSFPKSFSLPIQLES AVDTAKTSDVNSSV GK-STATLPLSFK EATLAKRFPALKTRS
m5	365	FAAAHRPKSQKCAVYKFR L VVKADGNQETNNGCHKVKIMP CFPFVAKEPS TKGLNPNP SH
m2	371	QPAKKKPP-PSRERK KVTRTI LAILLAFITWAFYNVHVLIN TFCAPGIPNVTWITIGYWLC
m4	383	NQVRKQRQMAARERK KVTRTI FAILLAFITWTPYNVHVLVNTFCQSGIPDITVWISIGYWLC
m1	348	QIAKRRKTFSLVKERK AARL SAILLAFITWTPYNIHVLVNTFCQSGIPKTLWESIGYWLC
m3	474	QITKRRKMSLVKERK AAGL SAILLAFITWTPYNIHVLVNTFCQSGIPKTLFWNLGYWLC
m5	425	QMTKRRKRVVLVKERK AAGL SAILLAFITWTPYNIHVLVNTFCQSGIPKTLVHLLGYWLC
m2	430	VINSTINPACYALCNATFKKTFKHLMLH YKNI GATR-----
m4	443	VVNSTINPACYALCNATFKKTFKHLMLH YRNI GATR-----
m1	408	VVNSTINPACYALCNATFKRDTFFRLLLLCR WDKRRWRK--IPKRP GSVHRTPSRQC-
m3	534	VVNSTINPACYALCNATFKRDTFFKMLLLLCQ CDKKKRRKQYVQRQSVI FHKRAPEQAL
m5	485	VVNSTINPACYALCNATFKRDTFFKMLLLLCR WKKKKVVEEKLYWQGN SKLP-----

13 loop



# Fungal Pheromone Receptors from Several Species (17-25% identical)

TMD1
TMD2

1 ---MLD HIT P P F F A L V A F F L V L M P F A W H I K S K N V G L I M L S I W L M L G N L D N F V N S M V W W K ---T T  
 1 M F S G K E N V S F G V L C L L A G C I S T S S C L I H L Q A K N I G V L L M M F W C F T G L V N K G I N A L A F N N ---S L  
 1 ---M S Y K S A I I G L C L L A V I L L A P P L A W H S H T K N I P A I I L I T W L L T M N L T C I V D A A I W S D D D F L T -  
 1 --M L P I G I F Y Q F Y A Y F A L V L S I P I L Y M Q L R A R N I P C L L L F W L T L T T L I Y V V E S A I W S N P Y A E T I

TMD3
TMD4

58 A D L A P A Y C E L S V R L R H L L F I A I P A S N L A I A R K L E S I I A S T R Q V R A G P G D H R R A V I I D L L I C L G I P I  
 62 R L A W T L G C D L S A I I E R T W O F G L C C S A L C V L Q R L E G I I A S L R Q A H S T V W D R K R R L L I D F G V G L G L P A  
 62 R W D G K G W C D I V I K L Q V G A N I G I S C A V T N I I Y N L H T I L K - A D S V L P D L S S W T K I V K D L V I S L F T P V  
 64 R W M G Y G L C D I T S R I V T C S S I G I P A S A F T L V L Y L D T V I R - R D H P L K R Y E N W - - - I W H V C L S I L L P L

TMD5
I3

123 I Y T S L M I V N O S N R Y G I L E E A G C W P M M V F S W L W V L L V A A P V I V V S L C S A V Y S A L A F R W F W V R R R Q F  
 127 L Q I P M F F I V Q P Y R L N V I E N I G C S A P L Y A S V P A L F I Y H L W R L L V S L V C A V Y A V L V L R W F M L R R R Q F  
 126 M V M G F S Y L L Q V F R Y G I A R Y N G C Q N L L S P T W I T T V L Y T M W M L I W S F V G A V Y A T L V L F V F P Y K K R K D V  
 125 I I M A M M V P L E S N R Y V V I C M N G C Y S S F Y Q T W Y T L L F F Y I P P C L L S F G G L F F V S R I V V L Y W R R Q R E L

TMD6

188 Q A V L A S S A S T I N R S H Y V R L L L L T A I D M L L F P P I Y V G T I A A Q I - K S S I S I P Y G S W S S V H T G F N Q I P  
 192 T A A L S S Q H S G L S Q K K Y F R L F A L A I C E R V L V S A G Q F Y V I I Q S L - Q I G G L L P Y T S W A E V H T N F N R I L  
 191 R D I L H C T N S G L N L T R F A R L L I E C F I I L V M F P F S V Y T F V Q D L Q Q V E G H Y T F K N T H S S T I W N T I I K  
 190 Q Q F F Q - R D S Q T T S K R F L R L T C L A A V F F L G Y F P L T I F M V V A N - G K L Q Q F L P F N H E L V E A W H Q E S T

TMD7

252 Q Y P A S L V L M E N T F Q R N L I L A R L V C P L S A Y I F F A M F G L G L E V R Q G Y K E A F H R A - L L F C R L R K E P K A  
 256 F V P V D T I A H S S L L - - S L S I L R W F S L T P A M A L F V F F G L T E E A Q S V Y K A R W K A L - I N L C - - - - S S K G  
 256 F D P G R P I - Y N I - - - - - - - - - - - W L Y V L M S Y L V F L I F G L G S D A L H M Y S K F L R S I K L G F V L D M W K R F I  
 253 Y Y P T T K V G L N D - - - - - - - - - - - W V P P T V L Y L M S L F E S T S G G W T E K V A L I L W S L L V W L P F T K - - - - -

316 S A L Q H V V A D I E Y V T F R S H D T F D A N T S T K S E K S D I D M R G S E A A - - - - - - - - - - - - - - - - -  
 314 K K Q T D G R E S L D L E A F E S H G - - - - - - - - - - - S K F S V L V Q R D T V I C - - - - - - - - - - - - - - - - -  
 310 D K N K E K R V G I L L N K L S S R K E S R N P F S T D S S E N Y I S T C T E N Y S P C V G T P I S Q A H F Y V D Y R I P D D P R K  
 303 N T A L G R H A Q F K L D C C K S I E S T M A G K T L D S T D F K E K C - - - - - - - - - - - L V L E R Q W S K S S I P S D N S S

- What is the range of sequence divergence among the sequences you plan to align?

```

enol1  EAMKMGAEVYHHLKSVIKKRYGQDATNVGDEGGFAPNIQENKEGL
enol2  EAMKMGCEVYHHLKAVIKKRYGQDATNVGDEGGFAPNIQENKEGL
enol3  EALRIGSEVYHNLKSLTKKRYGQSAGNVGDEGGVAPDIKTPKEAL
enol4  EAMKMGVEVYHNLKSIIKKRYGQDATNVGDEGGFAPNIQENKEGL
enol5  EALRIGSEVYHNLKSLTKKRYGASAGNVGDEGGVAPNIQTAEAL
enol6  EALKMGSEVYHALKSVIKAKRYGQDACNVGDEGGFAPNIQDNKEGL
enol7  EAMKMGSEVYHHLKKNVIKAKRYGLDATAVGDEGGFAPNIQSNKEAL
enol8  DAMRVGAEVYHSLKGVIKAKRYGKDATNVGDEGGFAPNILDNHEAL
cpeps  D-IEVADRVFTAAHRNVERRFGPVPLS-ASSGLMVPP--LDSAGQL

```

```

enol1  ELLKTAIAKAGYTGKVVIGMDVAASEFYG-SDKTYDLNFKKEENND
enol2  ELLKTAIEKAGYTGKVVIGMDVAASEFYG-KDKSYDLNFKKEESND
enol3  DLIMDAIDKAGYKGKVGIAMDVASSEFY--KDGKYDLDFKNPESD
enol4  ELLKAAIEKAGYTGKVVIGMDVAASEFFGEKDKTYDLNFKKEENND
enol5  DLIVDAIKAAGHDGKVKIGLDCASSEFF--KDGKYDLDFKNPNSD
enol6  ELLNEAIAKAGYTGKVKIGMDVASSEFY--KDGKYDLDFKNPNSD
enol7  NLISDAIAKAGYTGKIEIGMDVAASEFY--KDGQYDLDFKNEKSD
enol8  ELLKAAIAQAGYTDKVVIGMDVAASEFC--RDGRYDLDFKSP-PD
cpeps  DLLQAAVAETGHTEVCTLGVDVAA-EHLLTEPGRYRF-----

```

```

enol1  E A M K M G A E V Y H H L K S V I K K K Y G Q D A T N V G D E G G F A P N I Q E N K E G L
enol2  E A M K M G C E V Y H H L K A V I K K K Y G Q D A T N V G D E G G F A P N I Q E N K E G L
enol3  E A L R I G S E V Y H N L K S L T K K K Y G Q S A G N V G D E G G V A P D I K T P K E A L
enol4  E A M K M G V E V Y H N L K S I I K K K Y G Q D A T N V G D E G G F A P N I Q E N K E G L
enol5  E A L R I G S E V Y H N L K S L T K K R Y G A S A G N V G D E G G V A P N I Q T A E E A L
enol6  E A L K M G S E V Y H A L K S V I K A K Y G Q D A C N V G D E G G F A P N I Q D N K E G L
enol7  E A M K M G S E V Y H H L K N V I K A K F G L D A T A V G D E G G F A P N I Q S N K E A L
enol8  D A M R V G A E V Y H S L K G V I K A K Y G K D A T N V G D E G G F A P N I L D N H E A L
cpeps  D - I E V A D R V F T A A H R N V E R R F G P V P L S - A S S G L M V P - - L D S A G Q L

```

```

enol1  E L L K T A I A K A G Y T G K V V I G M D V A A S E F Y G - S D K T Y D L N F K E E N N D
enol2  E L L K T A I E K A G Y T G K V V I G M D V A A S E F Y G - K D K S Y D L N F K E E S N D
enol3  D L I M D A I D K A G Y K G K V G I A M D V A S S E F Y - - K D G K Y D L D F K N P E S D
enol4  E L L K A A I E K A G Y T G K V V I G M D V A A S E F F G E K D K T Y D L N F K E E N N D
enol5  D L I V D A I K A A G H D G K V K I G L D C A S S E F F - - K D G K Y D L D F K N P N S D
enol6  E L L N E A I A K A G Y T G K V K I G M D V A S S E F Y - - K D G K Y D L D F K N P N S D
enol7  N L I S D A I A K A G Y T G K I E I G M D V A A S E F Y - - K D G Q Y D L D F K N E K S D
enol8  E L L K A A I A Q A G Y T D K V V I G M D V A A S E F C - - R D G R Y D L D F K S P - P D
cpeps  D L L Q A A V A E T G H T E V C T L G V D V A A - E H L L T E P G R Y R F - - - - -

```

Enols 1-8: all >60% identical to each other  
Cpeps: <35% identical to Enols 1-8

gram.pos	L K A K - - G M N T A V <b>G D E G G Y</b> A P N L G <b>S</b> N D E <b>A L</b> A V I A
gram.neg	L S A K - - G M N T N V <b>G D E G G F</b> A P S L D <b>S</b> A S S <b>A L</b> D F I V
eukaryote	T K K R Y G A S A G N V <b>G D E G G V</b> A P N I Q T A E E <b>A L</b> D L I V
archaea.	L A D R - - D L P A G K <b>G D E G A W</b> A P S V - <b>S</b> D D E <b>A</b> F E I M D
<b>cpeps</b>	<b>V E R R F G P V P - - L S A S S G L M V P L D S A G Q - L D L L Q</b>

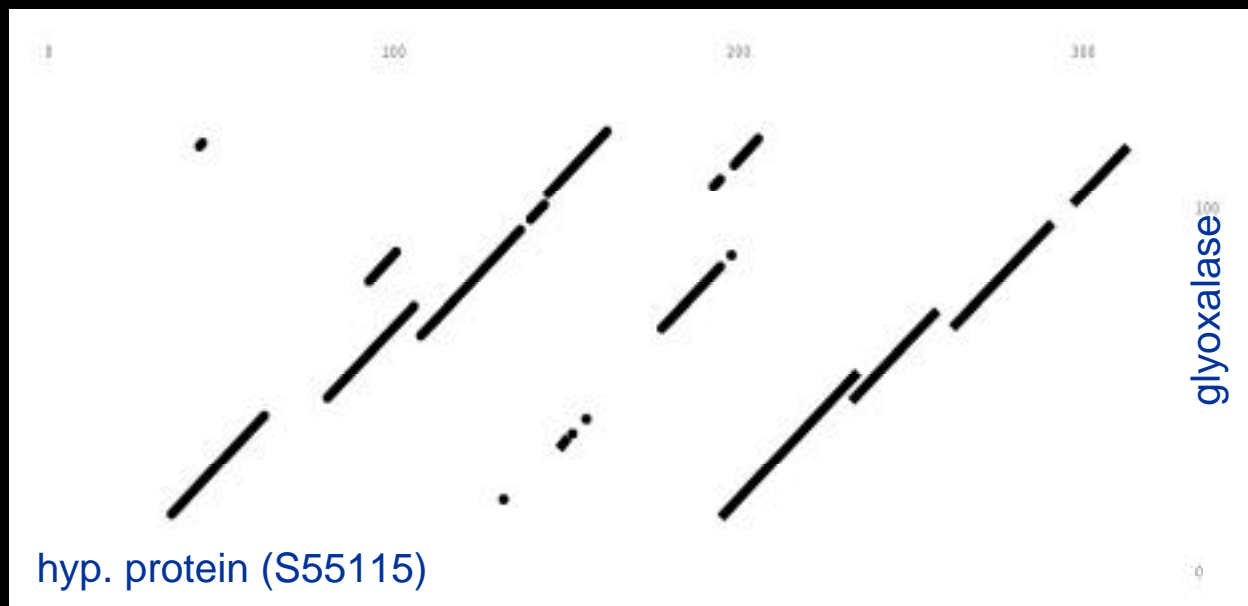
gram.pos	E <b>A</b> V K A A <b>G</b> Y E L G K D I T L A M <b>D C</b> A A S <b>E</b> F Y K D - G K - -
gram.neg	D S I S K A <b>G</b> Y K P G E D V F I A L <b>D A</b> A S S <b>E</b> F Y N K - D Q N I
eukaryote	D <b>A</b> I K A A <b>G</b> H - - D G K V K I G L <b>D C</b> A S S <b>E</b> F F K D - G K Y D
archaea.	E <b>A</b> V E T V A D D F G F A I S F G L <b>D V</b> A R A <b>E</b> L Y D D - E A D G
<b>cpeps</b>	<b>A A V A E T G H - - T E V C T L G V D V A A E H L L T E P G R Y R</b>

gram.pos	Y V L A - - - G E <b>G N</b> K A F <b>T</b> S E E F T H F L E E <b>L T</b> K Q <b>Y</b> P I V
gram.neg	Y D L K - - - G E <b>G R</b> K - L T S A Q L V D Y Y V E <b>L C</b> G K <b>Y</b> P I Y
eukaryote	L D F K N P N S D K S <b>K W L T</b> G P Q L A D L Y H S <b>L M</b> K R <b>Y</b> P I V
archaea.	Y V Y - - - - D D <b>G V</b> <b>K</b> - - S T E E Q I E Y I A G K V E E <b>Y</b> D L V
<b>cpeps</b>	<b>F - - - - - - - G D R V L T A P D F A D H L A D L A H R F R M S</b>

Enols: 37-62% identical to each other

Cpeps: 35-50% identical to Enols

# How do you handle internal repeats?



q09751.N    V E R S K R **E** G I **L E L T** Y **N F G T E K** K E G P V **Y** I **N**  
 s55115.N    P D V F S A H G V **L E L T** H **N W G T E K** N P D Y K I **N N**  
 q09751.C    - - - - - **E** G L **L E L T** H **N W G T E K** E S G P V **Y** H **N**  
 s55115.C    - - V F S C **E** S V **L E L T** H **N W G T E** N D P N F H **Y** H **N**  
 glyox.        - - - - - **E** A V I **E L T** Y **N W G** V D **K** - - - - - **Y** E L

q09751.N    **G N** T E P K R **G F** **G H I C** F T **V D** N I E S A **C** A Y L **E** -  
 s55115.N    **G N** E E P H R **G F** **G H I C** F S **V** S D I N K T **C** E E L **E** -  
 q09751.C    **G N** D G D E K **G Y** **G H** **V** C I S **V D** N I N A A **C** S K F **E** -  
 s55115.C    **G N** - S E P Q **G Y** **G H I C** I S **C** D D A G A L **C** K E I **E** V  
 glyox.        **G** T - - - - - A Y **G H I** A L S **V D** N A A E A **C** E K I R Q

q09751.N    - - S K G V S F K K K L S D **G** K M **K** H I A F - - - - -  
 s55115.N    - - S Q G V K F K K R L S E **G** R Q **K** D I A F - - - - -  
 q09751.C    - - A E G L P F K K K L T D **G** R M **K** D I A - - - F L L D  
 s55115.C    K Y G D K I Q W S P K F N Q **G** R M **K** N I A - - - F L K D  
 glyox.        N G G N V T R E A G P V K **G** - - - - - T T V I A F V E D

# General Issues in Multiple Alignment

- Computational complexity: a true multiple alignment of  $N$  sequences would require an  $N$ -dimensional matrix
- No single "correct" multiple alignment can be achieved except in trivial cases
- Methods assume sequences are independent rather than related by a phylogenetic tree in which the "branches" may evolve at different rates and with different positions being important to function

# Some Primary Algorithms for Multiple Alignment

- Global alignment methods construct an alignment throughout the length of the entire sequence
  - Examples: Pileup, Clustal family, MSA
- Local alignment methods identify ordered series of motifs, then aligns the intervening regions
  - Examples: MACAW, PIMA
- 1D profile analysis

## PILEUP (in GCG package\*)

- 1) Calculates a diagonal matrix of  $N(n-1)/2$  distances between all sequence pairs of  $N$  sequences using Needleman-Wunsch algorithm
- 2) Constructs a guide tree (dendrogram) from the distance matrix to direct the order of addition of subsequent pairwise alignments
- 3) Progressively aligns each cluster to the next most related sequence or cluster of sequences, adjusting the position of indels in all sequences

\*Genetics Computer Group, Madison, WI (available through UCSF SACS)

## Issues in the use of PILEUP

- Fast, generates reasonable alignments
- Current implementation in GCG handles up to 500 sequences
- All alignments determined from pairwise alignments, losing the information contained in the multiple alignment for position-specific scoring
- Overrepresentation of a subset of sequences to be aligned may bias the inference of an ordered series of motifs

# ClustalW\*

- From a family of programs using profile-based progressive alignment
- **Access:** <http://www2.ebi.ac.uk/clustalw/>
- Permits user adjustment of many parameters for both the pairwise and multiple alignment stages
- Computes position-specific gap opening and extension penalties as the alignment proceeds, *e.g.*, varies parameters at different positions

\*"W" stands for "weighting" the sequences to correct for unequal sampling of sequences from different evolutionary distances

## Steps in a ClustalW alignment

- 1) Constructs a distance matrix of all  $N(N-2)/2$  pairs using dynamic programming and converts scores to distances
- 2) Generates a "guide tree" using the neighbor-joining clustering algorithm of Saitou & Nei
- 3) Progressively aligns sequences in order of decreasing similarity using variable parameters and position-specific gap penalties

## The Bottom Line... \*

- For multiple alignments of divergent proteins, e.g., <30% identity, none of these methods is very satisfactory, suffering from 3 types of problems:
  - Inability to produce a single multiple alignment from correctly aligned subsets of the input sequences
  - Sensitivity to the number of sequences used
  - Sensitivity to the specific sequences used for multiple alignment

\*from the McClure paper listed in the lecture references

# 1-D Profile analysis

- Access: GCG package at SACS and at [http://www.sdsc.edu/projects/profile/new/help\\_main.html](http://www.sdsc.edu/projects/profile/new/help_main.html) (Gribskov, M., McLachlan, E.D., Eisenberg, D. (1987) *PNAS USA*, 84:4355-4358)
- Information in a multiple alignment is represented quantitatively as a table of position-specific symbol comparison values and gap penalties
- All information in the alignment is used
- Implementations available for both for database searching/sequence alignment

# Hidden Markov Models

- Probability-based models for database searching, multiple alignments, family generation (Pfam)
- Software and tools sites:
  - <http://hmmer.wustl.edu/>
  - <http://www.cse.ucsc.edu/research/compbio/HMM-apps/HMM-applications.html>
  - also at UCSF SACS

# Precomputed Multiple Alignments of Protein Families

- **Pfam:** <http://pfam.wustl.edu/>
  - Multiple sequence alignments and HMMs for many protein domains (3071 models as of 8/01)
- **Prodom:** <http://protein.toulouse.inra.fr/prodom.html>
  - Families generated automatically using PSI-BLAST with a profile built from the seed alignments of Pfam
- **Systems:** <http://www.dkfz-heidelberg.de/tbi/services/documentation/systemshelp.html>
  - Families clustered from SW-Prot/PIR using sequence walks and aligned via ClustalW
- **MetaFam:** <http://metafam.ahc.umn.edu/>
  - Functional assignments and a tool for comparison of how other family databases have made the classification

# Finding and Analyzing Motifs

# Applications for Motif Analysis

- Identification of very distant homologs
- May point to important functional units in a protein
- Can be used to "anchor" a multiple alignment
- Databases of motifs can be used to develop other informatics applications

Example: BLOCKS    Blosum

See: Bork, P. & Gibson, T. J. "Applying Motif and Profile Searches," in Methods in Enzymology 266: Computer methods for macromolecular sequence analysis, pp. 162-184

# Prosite: Protein Family Signatures

<http://tw.expasy.org/prosite/>

- Contains signatures for ~1500 families/domains
- Can be accessed using description, accession number, author, citation, full text search
- Provides several useful tools allowing a user to
  - Scan a sequence against a PROSITE pattern
  - Scan a pattern generated by a user or from PROSITE against the Swiss-Prot database
  - Scan a sequence against Profile databases, e.g., generalized profiles derived from multiple alignments
  - Many other specialized tools for motif/pattern generation and analysis
  - Includes substantial meta data: experts on each system, references, some statistical analysis

# Meme & Mast

<http://meme.sdsc.edu/meme/website/>

- **Meme: motif discovery tool**

(Grundy, W. M. et al. 1997. CABIOS 13, 397)

- motifs represented as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern
- output can be converted to BLOCKS which can then be converted to PSSMs (position-specific scoring matrices)

- Mast: database searching tool using one or more motifs as queries
  - provides a match score for each sequence in the database compared with each of the motifs in the group of motifs provided represented as P-values
  - provides probable order and spacing of occurrences of the motifs in the sequence hits

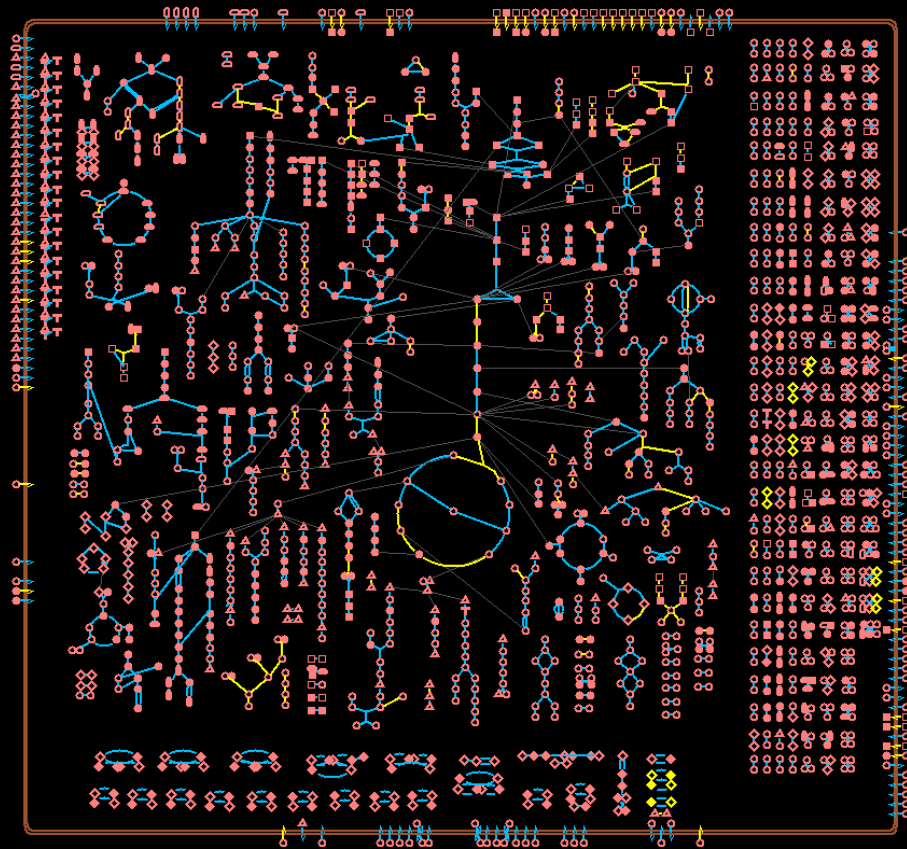
# **New Directions in Bioinformatics**

# Using Protein Informatics for Really New Insight into Biology

- **Comparative genomics**
  - **Metabolic computing: EcoCyc & MetaCyc**  
<http://ecocyc.org/ecocyc/index.html>
  - **Clusters of Orthologous Groups (COGS)**  
<http://www.ncbi.nlm.nih.gov/COG/>
- **Genetic circuits/Systems analysis**  
<http://gobi.lbl.gov/~aparkin/index.html>
- **Protein-Protein Interactions**
  - **Co-evolution**

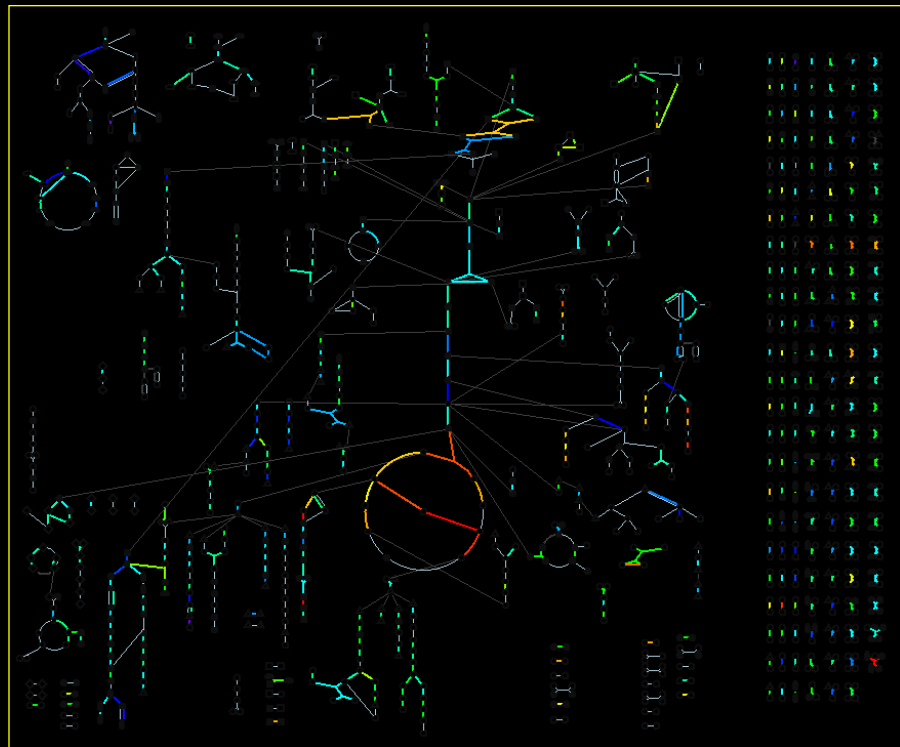
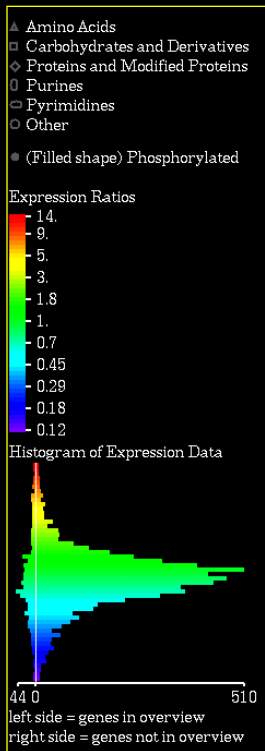
# Overview of *E. coli* metabolic systems

used with permission: Peter D. Karp (EcoCyc)



# MetaCyc: Yeast Expression Data

used with permission: Peter D. Karp (EcoCyc)



## A few important topics we didn't even mention

- Mapping Sequence Structure Function
- Structural superposition and 3D motif finding
- The 3D genome project
- Mapping the protein universe
- Census studies (Gerstein)
- Informatics for Proteomics
  - post-translational modifications
  - investigating protein machines

## See also:

- **Nucleic Acids Res. 2002 30**
  - Description and useful information on 112 databases of interest to the genomics/proteomics/bioinformatics communities