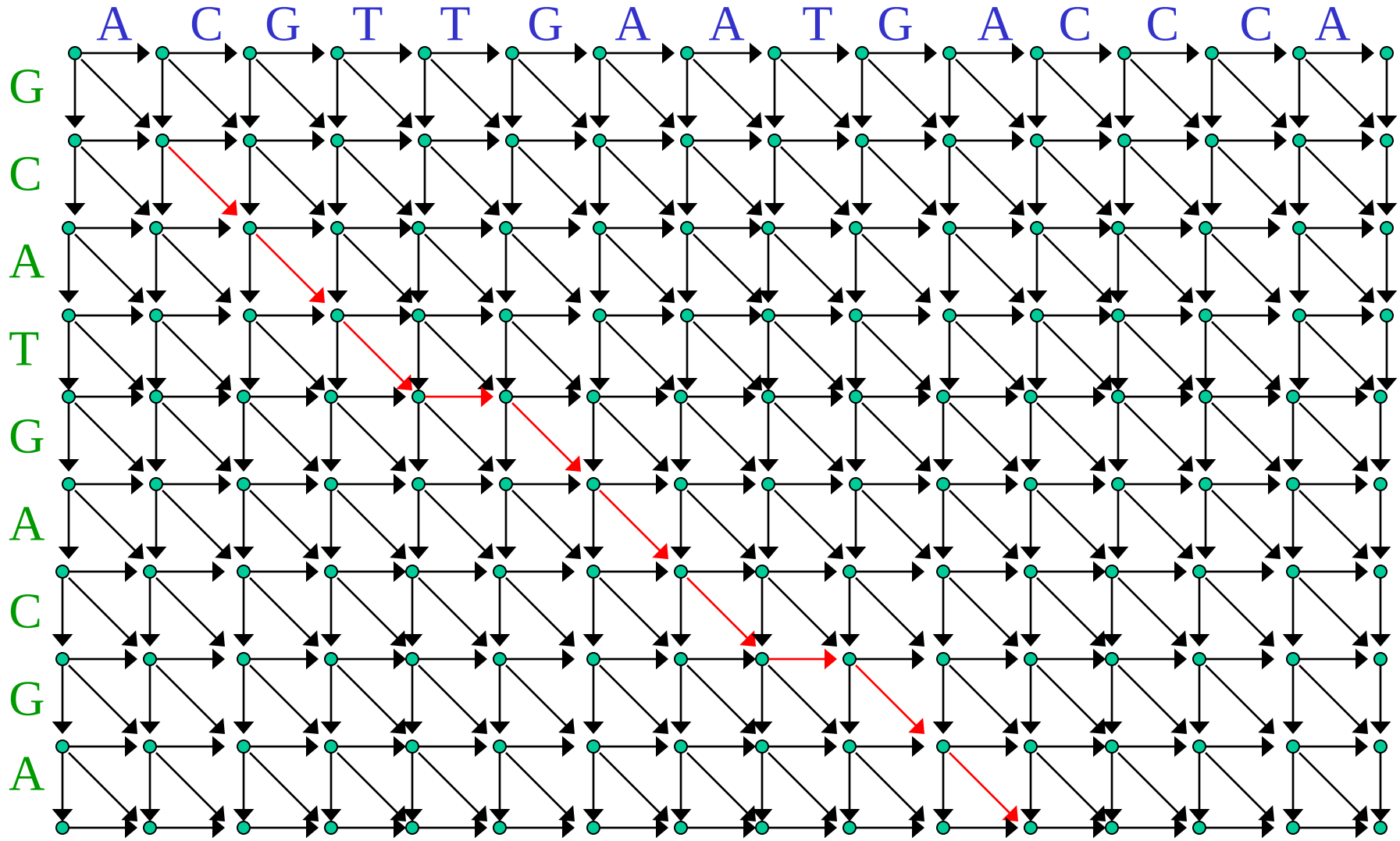
Lecture 10

- Local vs global alignments
- Alignment scoring
  - ‘Background’ models for proteins
    - Failure of equal frequency assumption
  - Score matrices
  - Profiles
- Statistical significance



Above path corresponds to following alignment (w/ lower case letters considered unaligned):

```
aCGTTGAATGAccca
gCAT-GAC-GA
```

Alignment algorithms

- *Smith-Waterman* algorithm to find highest scoring alignment

  = dynamic programming algorithm to find highest-weight path

  – is a *local* alignment algorithm:
    - finds alignment of subsequences rather than the full sequences.
- Can process nodes in any order in which parents precede children. Commonly used alternatives are
  – depth order
  – row order
  – column order

- If constrain path to
  - start at upper-left corner node and
  - extend to lower-right corner node,

  get a *global* alignment instead
- This sometimes called *Needleman-Wunsch algorithm*
  - (altho original N-W alg treated gaps differently)
- ∃ variants which constrain path to
  - start on the left or top boundary,
  - extend to the right or bottom boundary.

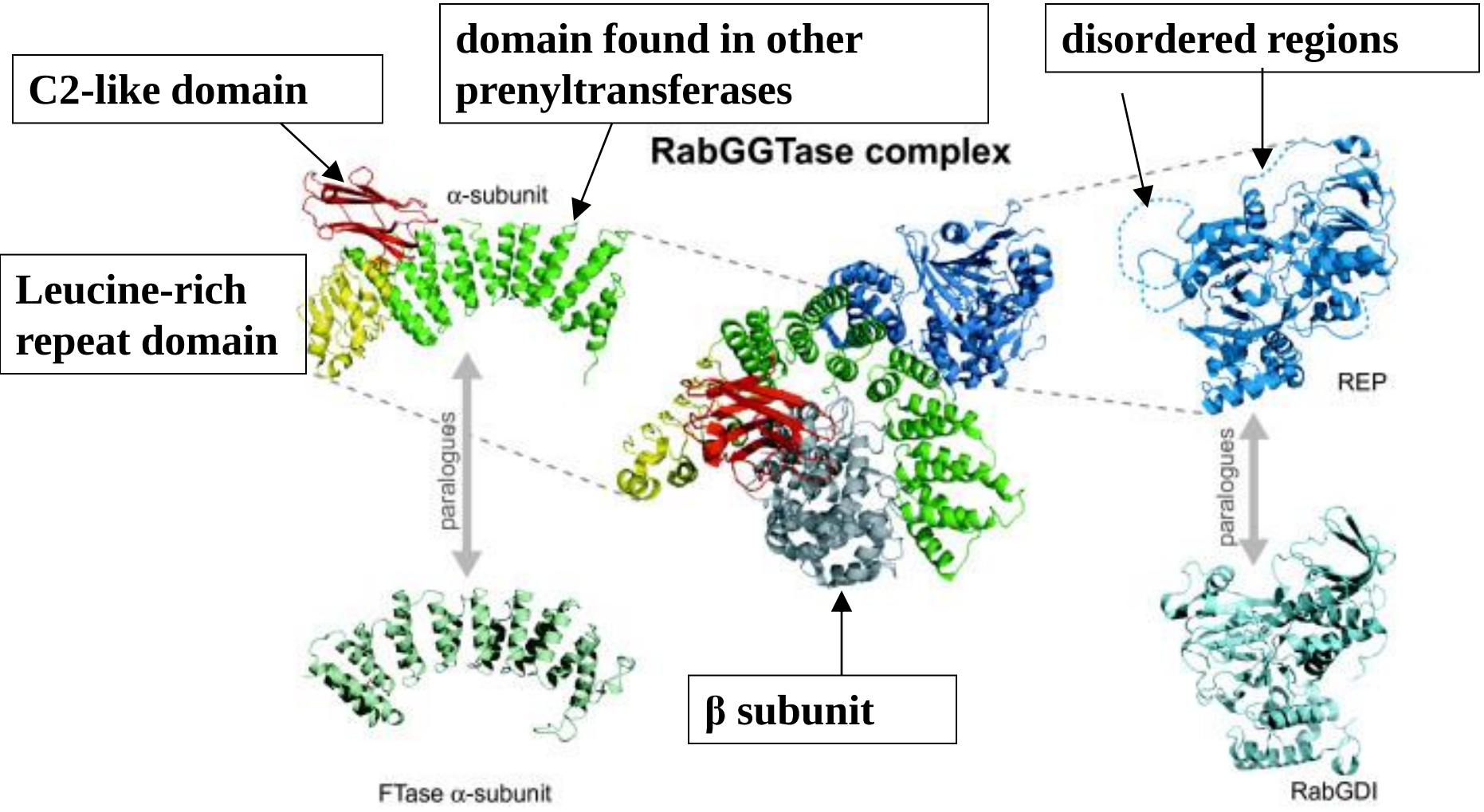
Local vs. Global Alignments: Biological Considerations

- Many proteins consist of multiple 'domains' (modules), some of which may be present
  - with similar, but not identical sequence

  in many other proteins
  - e.g. ATP binding domains, DNA binding domains, protein-protein interaction domains ...

  Need *local alignment* to detect presence of similar regions in otherwise dissimilar proteins.
- Other proteins consist of single domain evolving as a unit
  - e.g. many enzymes, globins.

  Global alignment sometimes best in such cases
  - ... but even here, some regions are more highly conserved (more slowly evolving) than others, and most sensitive similarity detection may be local alignment.
- Even local alignments can be misleading! (e.g. two nearby shared domains separated by non-homologous sequence)

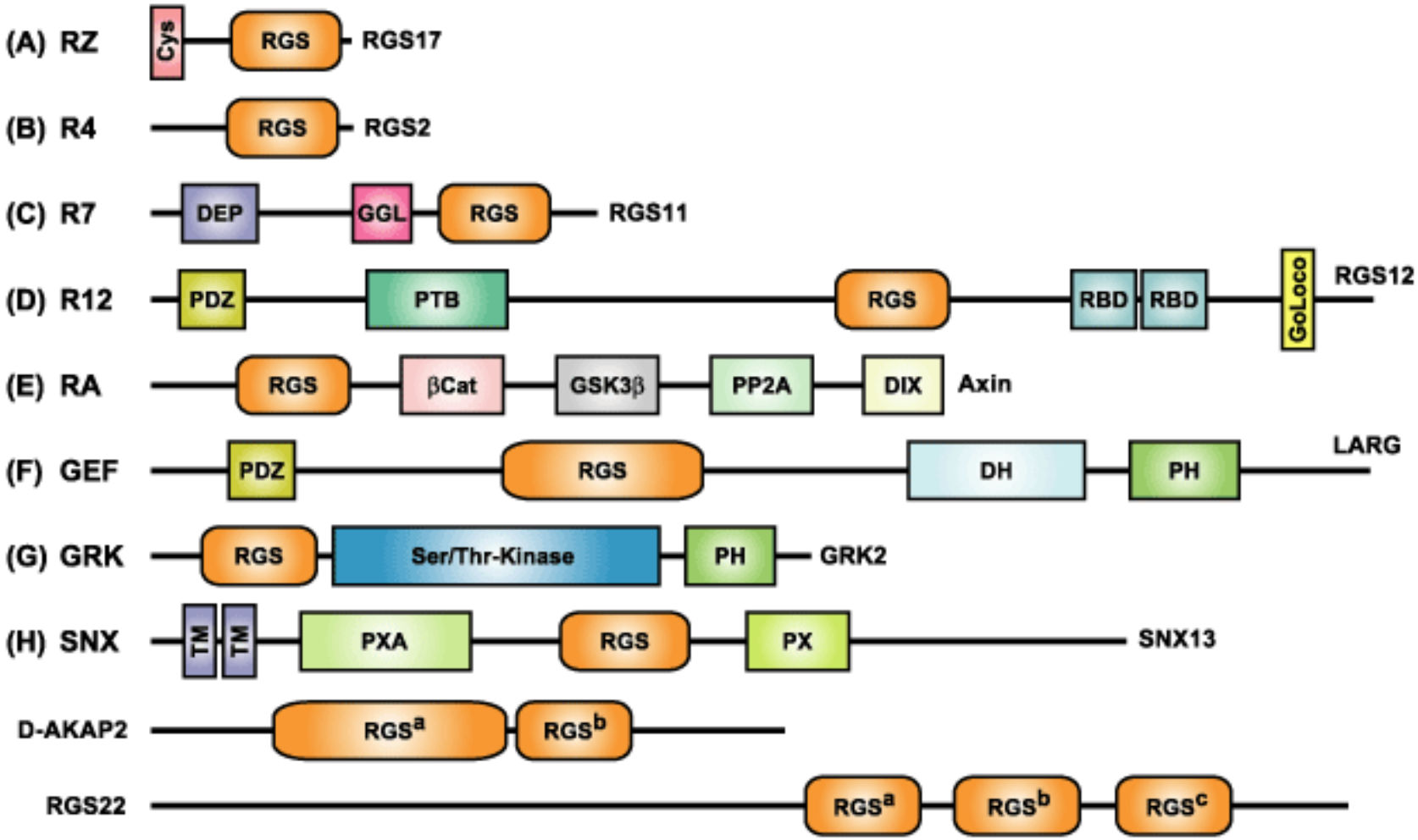


3-D structures of rat Rab Geranylgeranyl Transferase complexed with REP-1, + paralogs.

adapted from **Rasteiro and Pereira-Leal *BMC Evolutionary Biology* 2007 7:140**

Multidomain architecture of representative members from all subfamilies of the mammalian RGS protein superfamily.

from www.unc.edu/~dsiderov/page2.htm

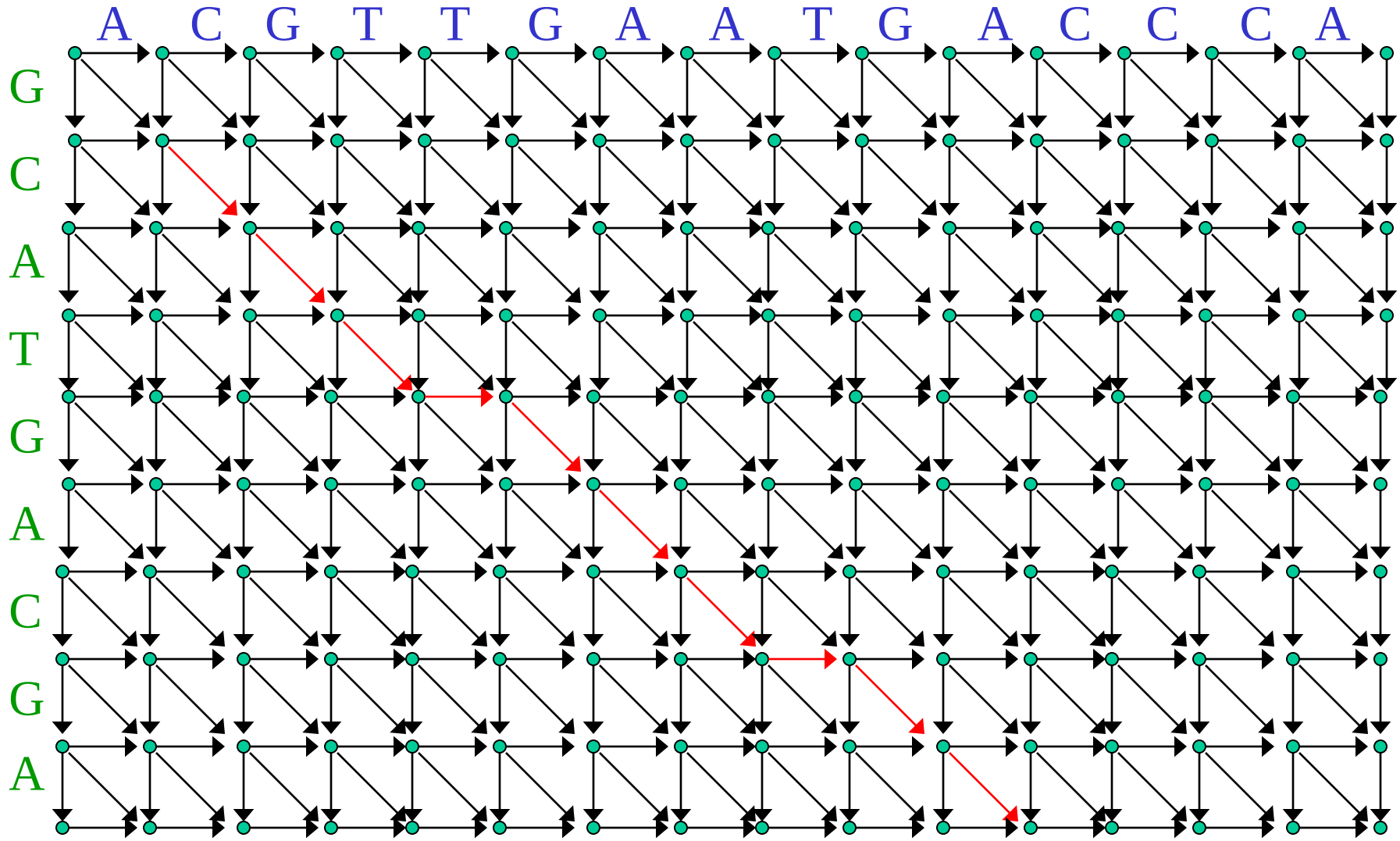


Similar considerations apply to aligning DNA sequences:

- (semi-)global alignment may be preferred for aligning
  - cDNA to genome
  - recently diverged genomic sequences (e.g. human / chimp)

  but local alignment often gives same result!
- between more highly diverged sequences, have
  - rearrangements (or large indels) in one sequence vs the other,
  - variable distribution of sequence conservation,

  & these usually make local alignments preferable.
- Genomic alignments are nearly always done in ‘chunks’



Above path corresponds to following alignment (w/ lower case letters considered unaligned):

```
aCGTTGAATGAccca
gCAT-GAC-GA
```

Weights on Edit Graphs

- Edge weights correspond to scores on alignment columns.
- Highest weight path corresponds to highest-scoring alignment for that scoring system.
- Weights may be assigned using
  - a *substitution score matrix*
    - assigns a score to each possible pair of residues occurring as alignment column
  - or *profile*
    - scores specific to a particular sequence

  and
  - a *gap penalty*
    - assigns a score to column consisting of residue opposite a gap.
  - Example for protein sequences: BLOSUM62

Alignment Scoring

- Optimal alignment scoring depends on probabilistic modelling (e.g. LLR scores)

- Default approach:
  1. each alignment column (edge in WDAG) is scored independently
     $\longrightarrow$ an independence assumption for probability model
  2. Score depends only on the residues that are present (via a BLOSUM-type score matrix) – i.e. independently of position within sequence

‘Background’ models for *protein* sequences

- The *independence* assumption is (usually) OK
  - Main violation: low complexity regions

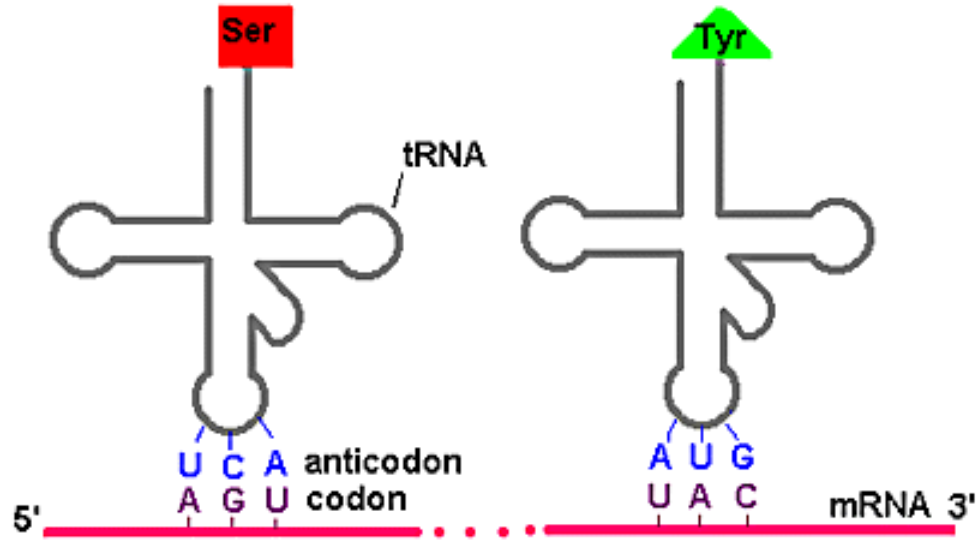
- The *equal frequency* assumption is not

Failure of equal frequency assumption for proteins

AMINO ACID	FREQUENCY.	# SYNON CODONS.
L	.093	6
A	.075	4
S	.072	6
G	.069	4
V	.065	4
E	.063	2
K	.059	2
T	.058	4
I	.057	3
D	.053	2
R	.052	6
P	.049	4
N	.045	2
F	.041	2
Q	.040	2
Y	.032	2
M	.024	1
H	.022	2
C	.017	2
W	.013	1

Hypotheses to explain correlation between frequency and # codons

- (*Neutralist*):
  - Nucleotide sequences that encode proteins are on average close to random,
  - so amino acid freqs are proportionate to codon freqs in random DNA.
- (*Selectionist*):
  - The genetic code evolved concurrently with early proteins, and
  - is adapted so that the most useful amino acids are encoded by the most codons.
- The truth is probably some combination of these!
  - Dependence of aa composition on genomic G+C content is consistent with neutralist hypothesis



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code

Deviations from ‘randomness’

- Compute, for each residue *r*, the ratio obs_r / exp_r of
  - the observed frequency obs_r, to
  - the expected frequency exp_r if coding sequences were random:

$$exp_r = (\#\text{codons encoding } r) / 61$$

Amino Acid	Obs/Exp	1st codon base	2nd codon base	3rd codon base	# codons
E	1.92	G	A	R	2
K	1.80	A	A	R	2
D	1.62	G	A	Y	2
M	1.46	A	T	G	1
N	1.37	A	A	Y	2
F	1.25	T	T	Y	2
Q	1.22	C	A	R	2
I	1.16	A	T	Not G	3
A	1.14	G	C	N	4
G	1.05	G	G	N	4
V	.99	G	T	N	4
Y	.98	T	A	Y	2
L	.95	C(T)	T	N	6
T	.88	A	C	N	4
W	.79	T	G	G	1
P	.74	C	C	N	4
S	.73	T(A)	C(G)	N	6
H	.67	C	A	Y	2
R	.53	C(A)	G	N	6
C	.52	T	G	Y	2

Obs/Exp Ratios

- All observed values are within factor of 2 of expected;
  - last column suggests trend towards “correcting” disparate # codons
- At codon position 1,
  - purines (*A* and *G*) predominate among over-represented amino acids,
  - pyrimidines (*C* and *T*) among under-represented amino acids.
- At codon position 2,
  - *A* and *T* predominate among over-represented amino acids,
  - *C* and *G* among under-represented amino acids.
- Hypotheses to explain *RWR* codon preference:
  - (Neutralist) Vestige of ancestral code? (Shepherd)
  - (Selectionist) More efficiently translated?

BLOSUM62 Score Matrix

GAP -12 -2

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

- Matrix entries are of form

  $M(r, s) = \log_a(h_{r,s} / b_{r,s})$ (rounded to int) where

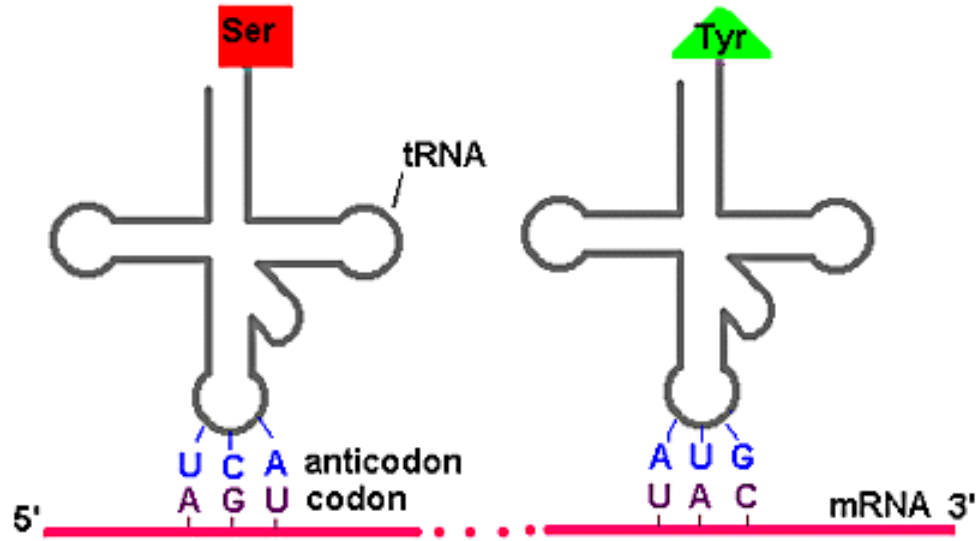
  $h_{r,s}$ = freq of $\begin{matrix} r \\ s \end{matrix}$ in homologous* seq alignments

  * '62' refers to specific set of homologue alignments

  $b_{r,s}$ = freq of $\begin{matrix} r \\ s \end{matrix}$ in 'background' (random) alignments

  a (the logarithm base) = $\sqrt{2}$ ('half bits')

- amino acid pairs with positive scores tend to be
  - *chemically similar*
  - in *same row or col* of genetic code table



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code

Improved scoring methods

- Ways to allow ***partial*** non-independence while preserving dynamic programming framework:

  1. Allow scores to depend on position within the sequence
     - so some substitutions (of same residues) or gaps penalized more heavily than others
     - like a site model!
  2. Enhance graph
     - Allows ‘memory’ of preceding columns

Profiles (position-specific scoring)

- Different parts of sequence may evolve at different rates
- In proteins
  - conserved functional motifs
  - structural constraints:
    - internal core region of tightly packed residues are more highly conserved;
    - surface residues, particularly in loops, often less conserved.

Conserved Domain in RecR and Class I Topisomerases

```
RecR   RLAEEKITEVILATNPTVEGEATANYIAELC
RecM   RLQDDQVTEVILATNPNIEGEATAMYISRLL
RecR   RVDDVGITEVIIATDPNTEGEATATYLVRMV
TrsI   IFKENKIDEVIIATDPARGGENIAYKILNQL
TOP1   KQLAEKADHIYLATDLDREGEAIAWRLREVI
ORF1   AELLKQANTIIVATDSDREGENIAWSIIHKA
TOP1   KDALKDADELILATDEDREGKVISWHLLQLL
TOP1   TIFDKRVKTIILATDAAAEGEYIGRNILYRL
TOP3   KREARNADYLMIWTDCDREGEYIGWEIWQEA
TOP3   KRFLHEASEIVHAGDPDREGQLLVDEVLDYL
RGYR   RNLAVEADEVLIGTDPDTEGEKIAWDLYLAL
```

CONSENSUS **xxxxxxxxxxU&uatDxxxEGexxxxxUxxxu**

Consensus key:

Uppercase: all residues chemically similar

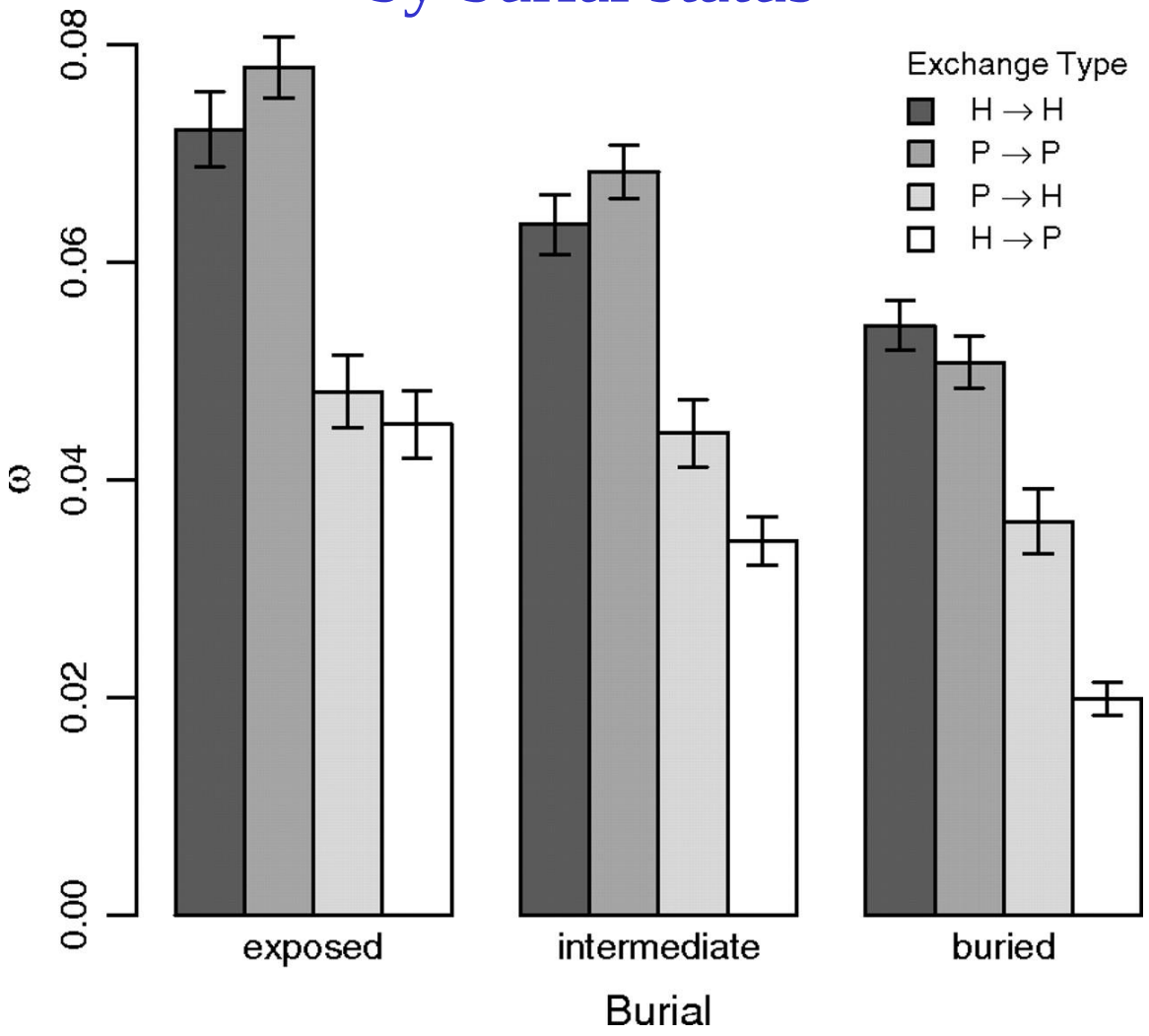
lowercase: most are

U,u: bulky aliphatic (I,L,V)

&: bulky hydrophobic (I,L,V,M,F,Y,W)

From RL Tatusov, SF Altschul, and EV Koonin, PNAS 91: 12091-12095

Rates of amino acid exchange in mammalian proteins by burial status



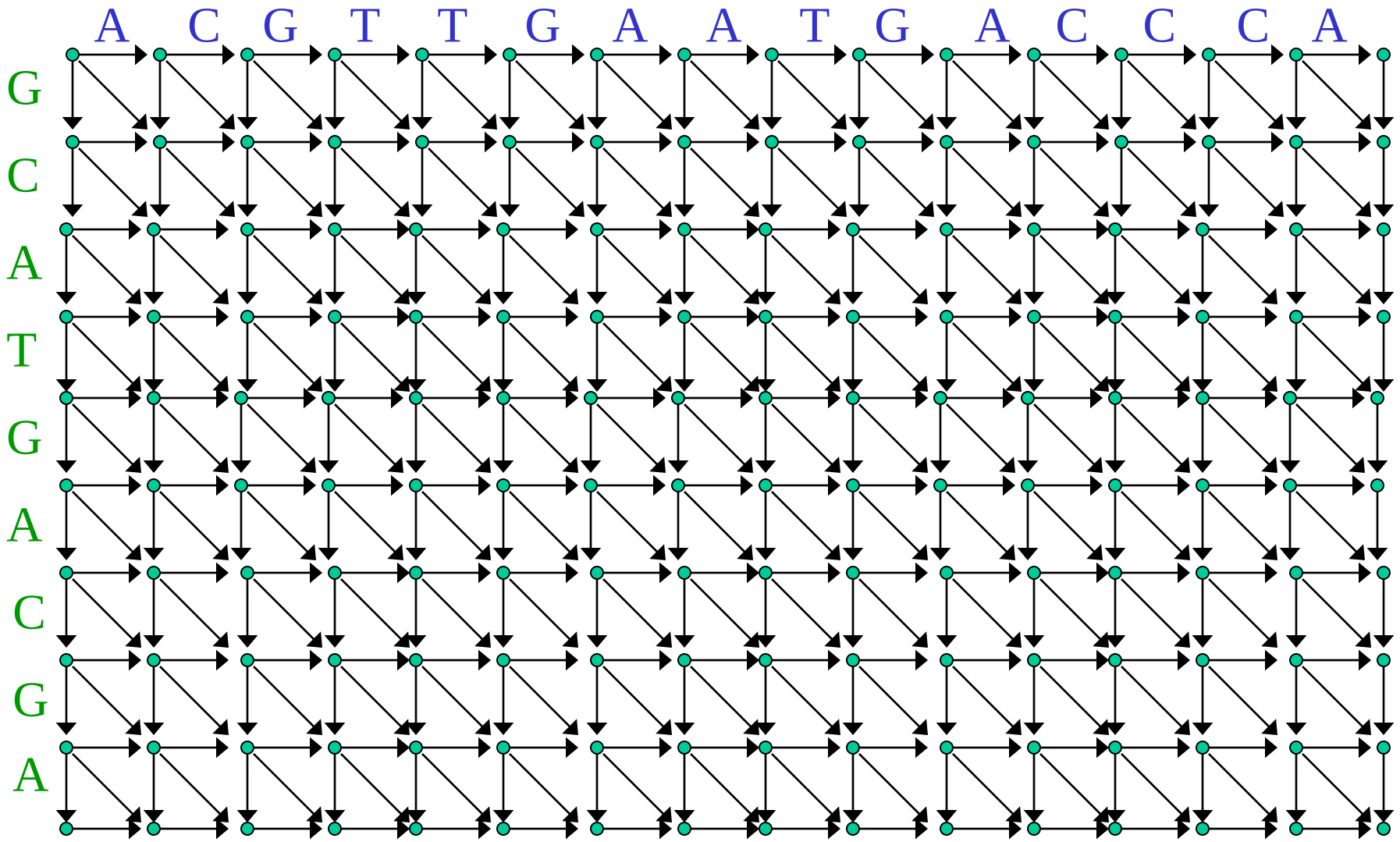
H: hydrophobic

P: polar

Saunders & Green Mol Biol Evol 2007 24:2632-2647; doi:10.1093/molbev/msm190

Molecular Biology and Evolution

The *Edit Graph* for a Pair of Sequences



- *Profiles*: *Position-specific* scoring scheme specifying score of each possible substitution at each position of a sequence

C →

Cons	A	C	D	E	...	T	V	W	Y	Open	Ext
G	7	-14	-1	-5	...	6	4	-34	-22	28	28
P	5	-26	4	1	...	1	-4	-48	-31	28	28
L	-18	-31	-40	-35	...	-16	13	-31	-9	100	100
T	7	-21	-4	-6	...	10	-3	-38	-28	100	100
E	6	-37	11	12	...	2	-10	-61	-38	100	100
A	5	-34	3	4	...	1	-8	-48	-34	100	100
E	0	-53	26	31	...	-5	-29	-60	-42	100	100
R	-11	-45	-11	-13	...	-3	-21	-2	-33	100	100
T	4	-28	-2	-1	...	8	7	-51	-24	100	100
M	-7	-47	-6	-6	...	-3	-6	-35	-26	100	100
V	0	-20	-22	-36	...	2	41	-56	-27	100	100
K	-9	-44	-11	-11	...	0	-5	-29	-31	100	100
N	5	-27	7	6	...	8	-11	-40	-32	100	100
A	7	-27	-4	-6	...	4	5	-46	-31	100	100
W	-47	-69	-58	-60	...	-40	-49	139	-6	100	100
G	11	-31	5	1	...	3	-5	-65	-43	100	100
K	-2	-46	5	8	...	-1	-23	-49	-45	100	100
V	-4	-23	-27	-45	...	-2	34	-48	-18	100	100
L	-3	-9	-6	-5	...	-3	3	-3	-1	26	26
N	-4	-26	3	2	...	-4	-19	-31	-9	26	26
A	4	-16	0	1	...	2	-12	-40	-10	26	26
H	0	-30	14	10	...	3	-15	-41	-21	100	100
I	-2	-20	-18	-23	...	-1	17	-50	-11	100	100
....											

***From* R. Luthy, I. Xenarios and P. Bucher, Improving the sensitivity of the sequence profile method *Protein Sci.* 3: 139-146 (1994)**

- The *scores* are ***position-specific*** LLRs (like a site model!):
- *Instead of*

  $M(r, s) = \log_a(h_{r,s} / b_{r,s})$ where

  $h_{r,s}$ = freq of $\begin{matrix} r \\ s \end{matrix}$ in homologous seq alignments

  $b_{r,s}$ = freq of $\begin{matrix} r \\ s \end{matrix}$ in 'background' (random) alignments
- *take, for i-th row* (*with residue* r_i)
  - $M_i(s) = \log_a(h_{i,s} / b_{i,s})$ where

    $h_{i,s}$ = freq of s aligned to r_i in homologue alignments

    $b_{i,s}$ = freq of s in random alignments

- PSIBLAST approach:
  1. initially compare query sequence to database sequences (using BLOSUM-type scoring matrix),
  2. build profile using matches
  3. rescan database using profile
  4. iterate 2 & 3 until …

Karlin / Altschul for sequence alignments

- For LLR-based alignment scoring
  - *i.e.* $s(r) = \log_a(t_r / b_r)$, where r is an alignment column,

the expected # local alignments of score $\geq S$ for (random) seqs of length M , N is

$$\approx MNK\, a^{-S}$$

for some constant K (not depending on S)

- Note that $a^{-S} = a^{-LLR} = 1 / LR$
- K-A developed theory for *ungapped* alignments, but empirical studies suggest it applies more broadly
  - Estimate K from alignments to random sequence