Structural Bioinformatics Lecture 6

Advanced sequence similarity search using profile HMMs.

Implications for protein evolution

UNIVERSITÄT DES SAARLANDES

Outline

- Alignment using Hidden Markov Models
- Profile alignment
- Profile alignment using Hidden Markov Models
  - HHsearch
- Evolution of proteins and the origin of life

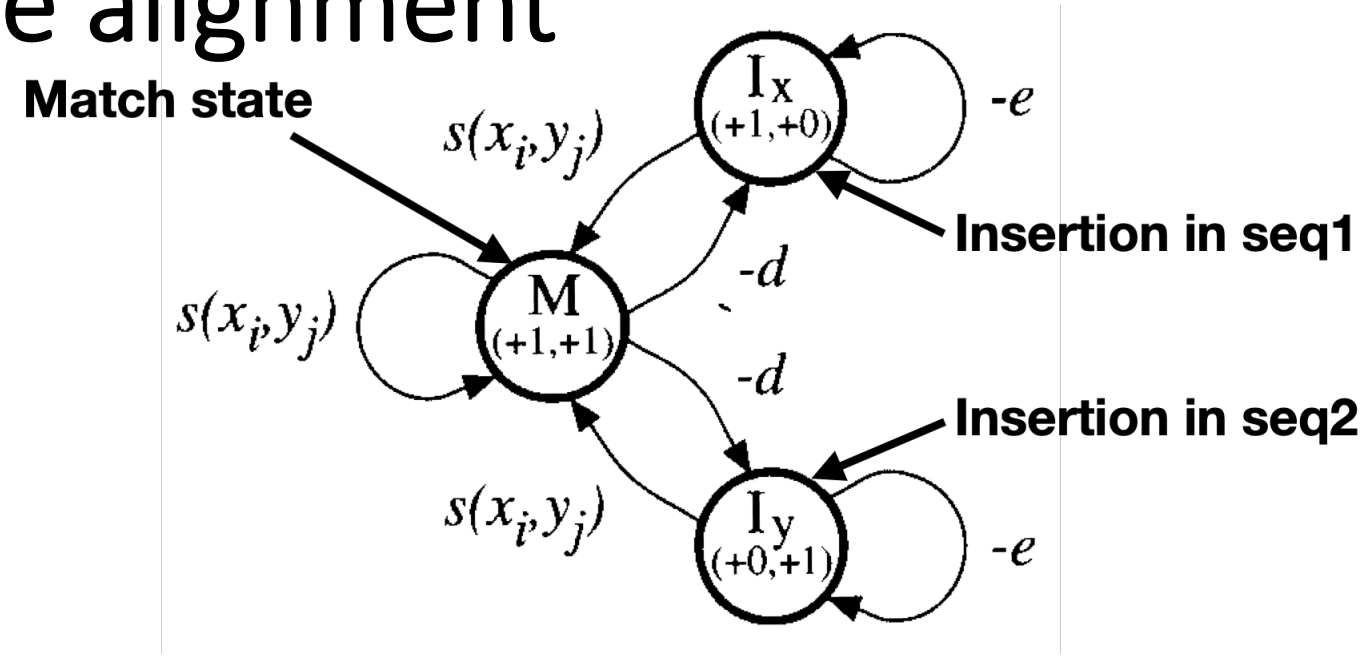
Alignment using HMMs

Based on:
Durbin et al., “Biological Sequence Analysis”, Cambridge University Press, 1998

Previously (lecture 4)…

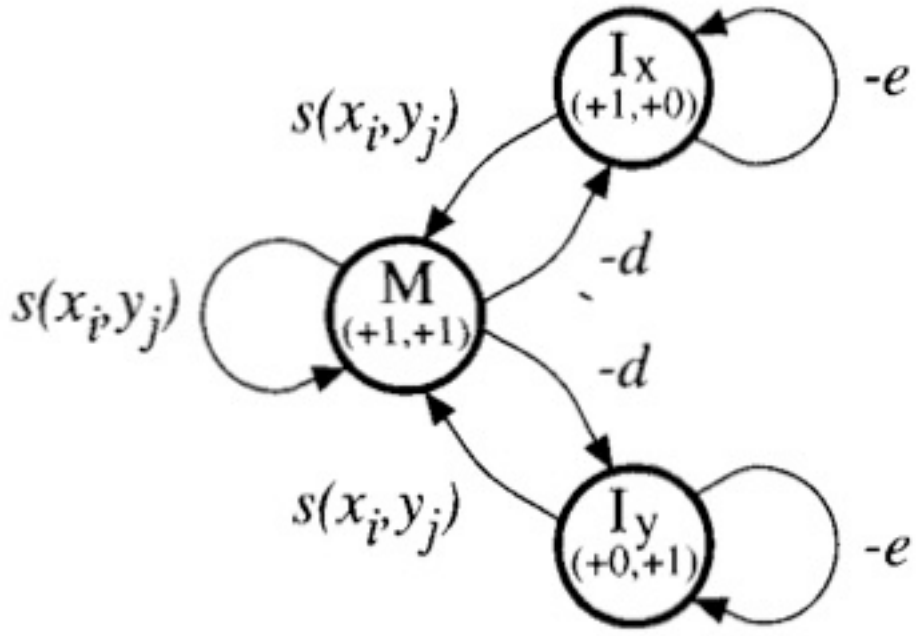
- Hidden Markov models for predicting TM segments
  - (and other stuff)
- Alignment of two sequences using dynamic programming
  - Can be implemented using other CS techniques

Finite state automaton (FSA) for pairwise sequence alignment



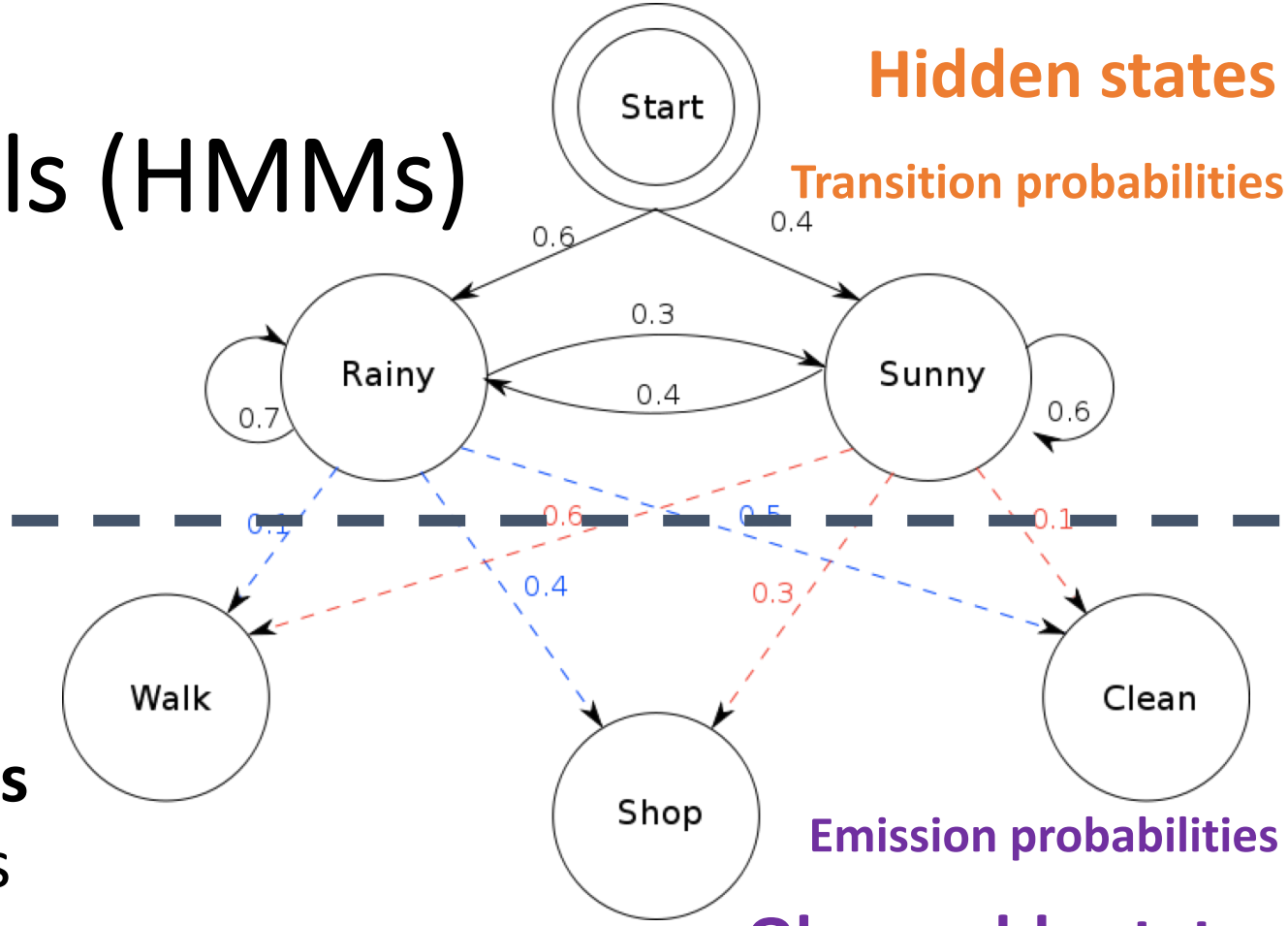
- **Process**: constructing an alignment pair by pair while traversing the two sequences
- Affine gap model naturally implemented: penalty for a gap of length g: $\gamma(g) = -d - e(g-1)$

Finite state automaton (FSA) for pairwise sequence alignment



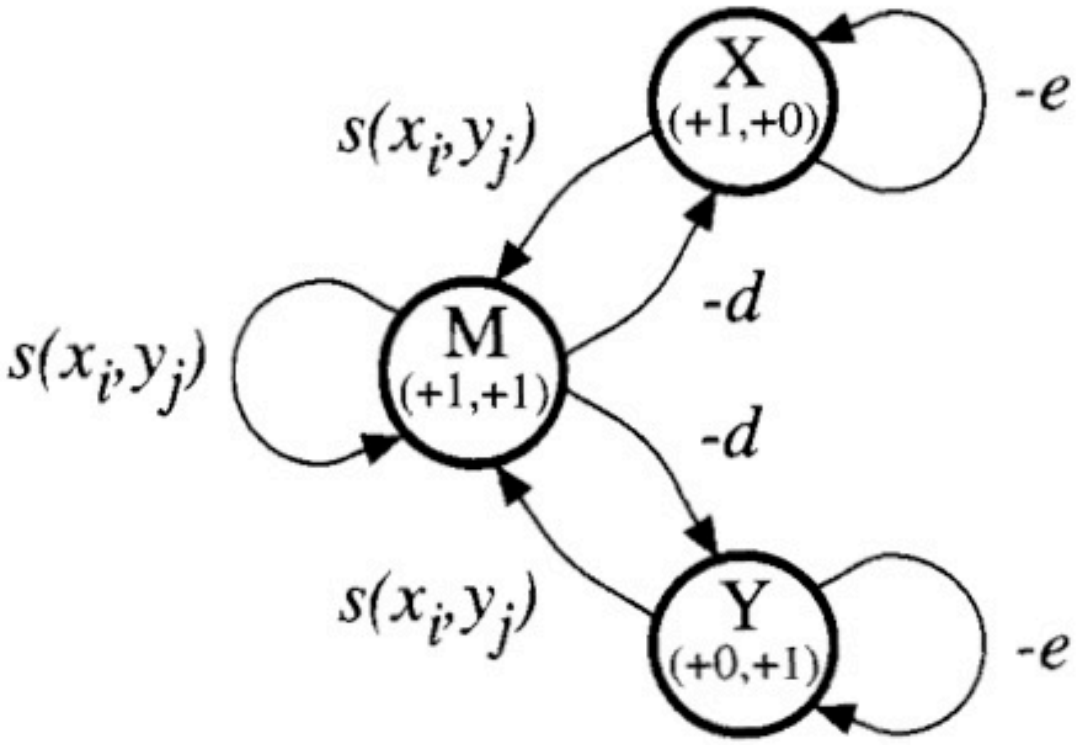
- Optimal path through the model can be reconstructed in **polynomial** time (see Durbin et al. for details)

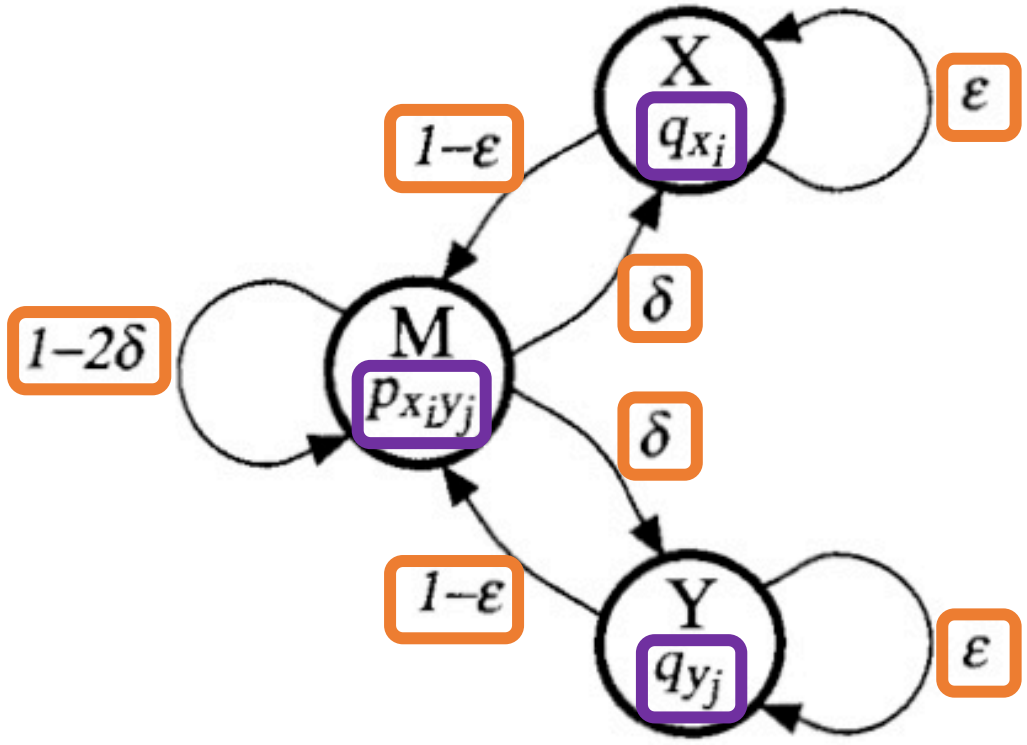
Hidden Markov models (HMMs)



- The modelled system is assumed to be a **Markov process** over unobserved (**hidden**) states
- Hidden states emit observable states
- **Task**: given the sequence of observable states, **reconstruct the most probable sequence of the hidden states**

Probabilistic version of the alignment FSA: an HMM

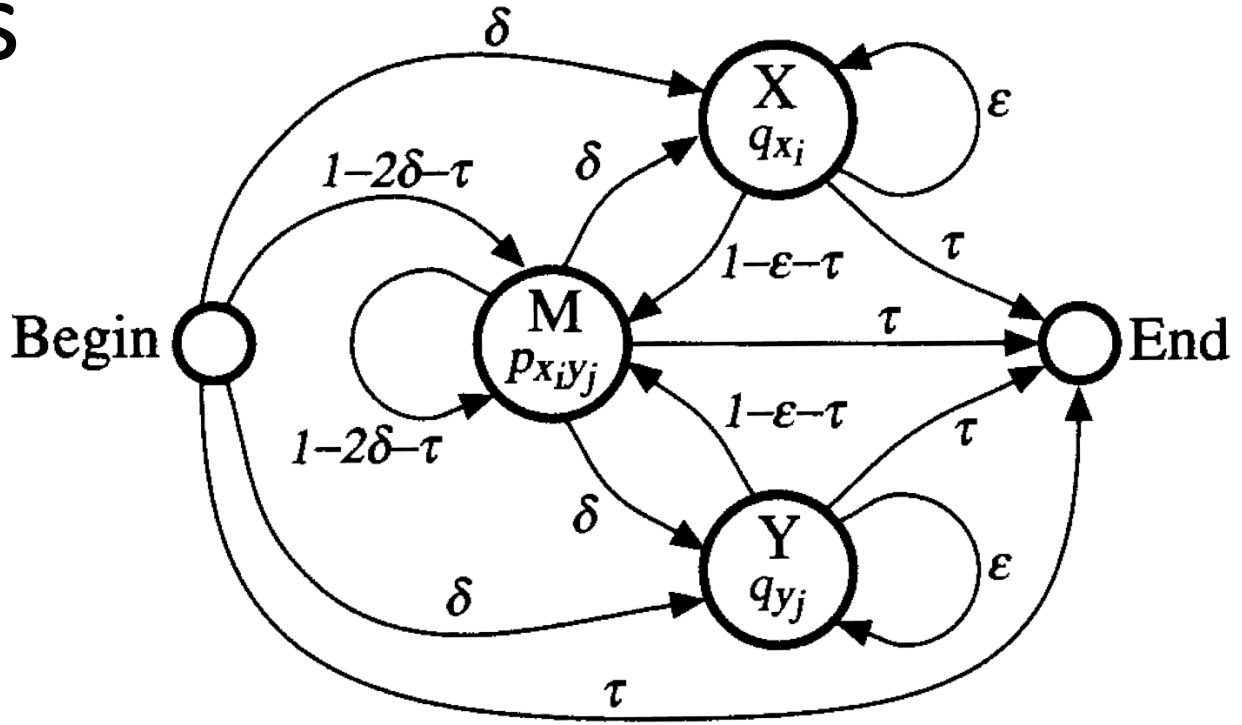




Transition probabilities

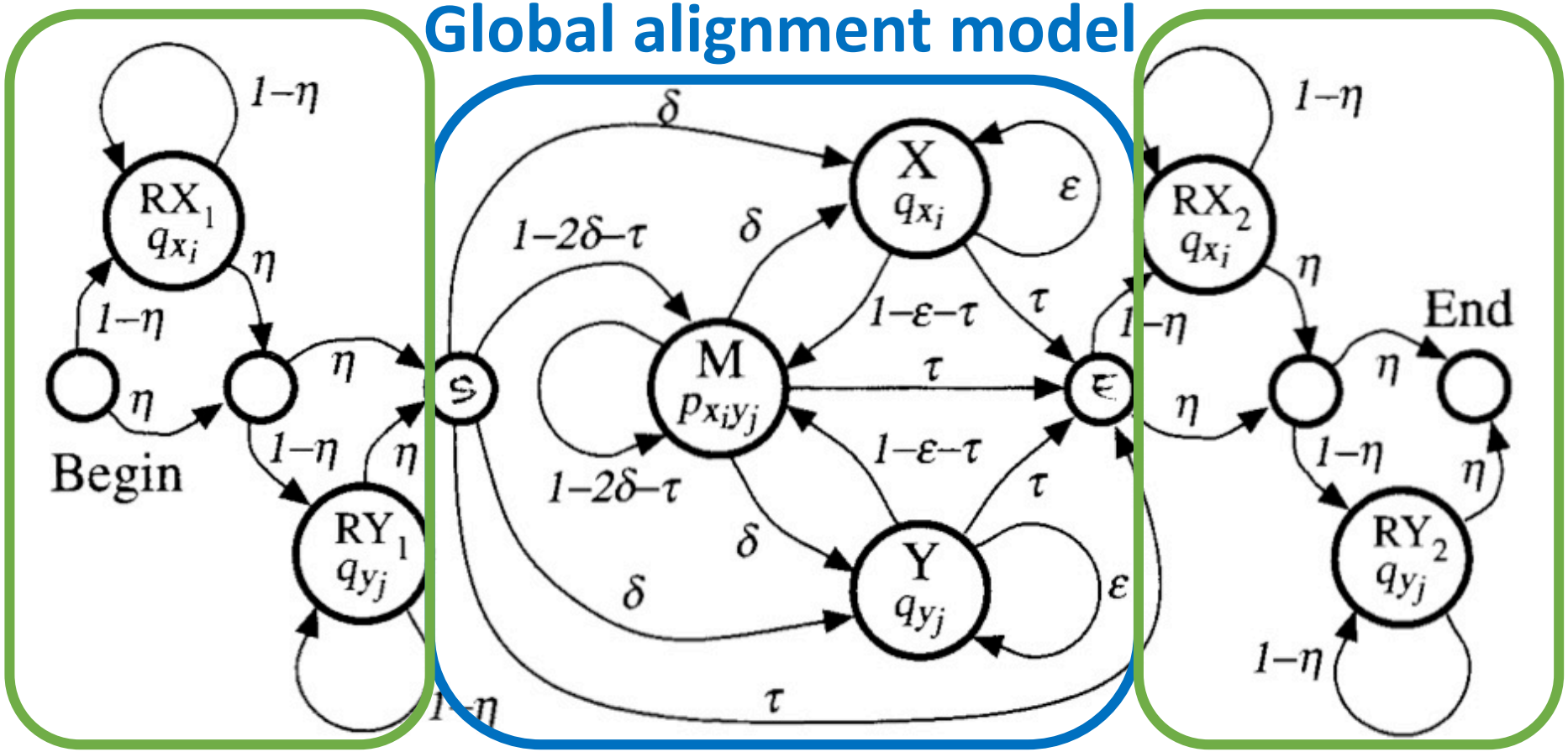
Emission probabilities

Full HMM for global alignment of two sequences



- Generates an aligned pair of two sequences
- Aligning with HMM: given two sequences *x* and *y*, find the most probable path through the model (e.g. Viterbi algorithm)

HMM for local alignment of two sequences



Random model: generates two non-aligned sequences

Random model: generates two non-aligned sequences

Profile HMMs for protein families

Based on:
Durbin et al., “Biological Sequence Analysis”, Cambridge University Press, 1998

Protein family profile

- Suppose we have collected a bunch of related proteins (e.g. with BLAST) and built an alignment (multiple sequence alignment methods will not be covered in this course)
  - Can we use it to find other related proteins?
  - Is this better (i.e. more sensitive) than BLASTing individually with the protein sequences?

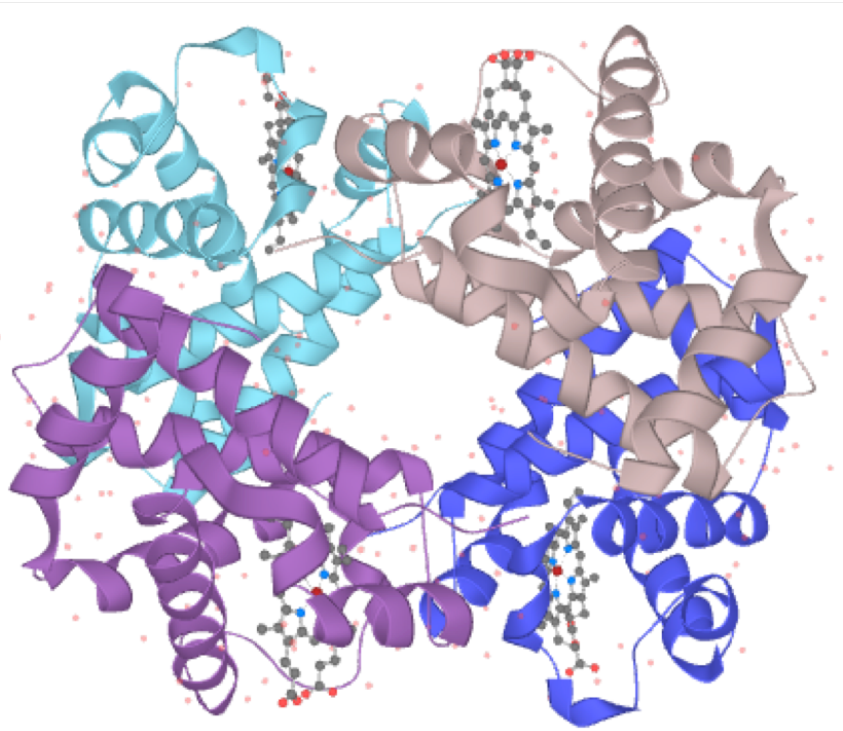
- YES — protein family **profile**

Conservation in MSA

```
Helix                 AAAAAAAAAAAAAAAA  BBBBBBBBBBBBBBBBB CCCCCCCCCCC
HBA_HUMAN  ---------VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF
HBB_HUMAN  ---------VHLTPEEKSAVTALWGKV---NVDEVGGEALGRLLVVYPWTQRFFESF
MYG_PHYCA  ---------VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRF
GLB3_CHITP ----------LSADQISTVQASFDKVKG-----DPVGILYAVFKADPSIMAKFTQF
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKF
LGB2_LUPLU --------GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-F
GLB1_GLYDI ---------GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-F
Consensus             Ls....  v a W kv .  .    g . L.. f . P .   F  F

Helix          DDDDDDDEEEEEEEEEEEEEEEEEEE          FFFFFFFFFFFF
HBA_HUMAN  -DLS-----HGSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL-
HBB_HUMAN  GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL-
MYG_PHYCA  KHLKTEAEMKASEDLKKHGVTVLTALGAILKK---K-GHHEAELKPLAQSHATKH-
GLB3_CHITP AG-KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG-
GLB5_PETMA KGLTTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF-
LGB2_LUPLU LK-GTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-
GLB1_GLYDI SG---AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGN
Consensus   .  t    ..  . v..Hg kv. a   a...l   d   . a l. l   H  .

Helix       FFGGGGGGGGGGGGGGGGGG    HHHHHHHHHHHHHHHHHHHHHHHHHH
HBA_HUMAN  -RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR-------
HBB_HUMAN  -HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-------
MYG_PHYCA  -KIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB3_CHITP --VTHDQLNNFRAGFVSYMKAHT--DFA-GAEAAWGATLDTFFGMIFSKM-------
GLB5_PETMA -QVDPQYFKVLAAVIADTVAAG----F----DAGFEKLMSMICILLRSAY-------
LGB2_LUPLU --VADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
GLB1_GLYDI KHIKAQYFEPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGALISGLQS-----
Consensus   v.   f  l . ..  ....    f   . aa. k. .    l sky
```



Helices are better conserved than loops between them!

Protein family profile: idea

- Some protein regions are better conserved than other
  - Conserved sequence regions often correspond to conserved / structured 3D regions
- Some residues are particularly conserved
  - They often correspond to functionally important residues
- Conserved regions and residues are more important
- **When searching / aligning new members of a family, it makes sense to concentrate on these conserved regions and residues**

Profile HMM

- **Ungapped regions**: probabilities of amino acids at each position are different
- => Specify probabilities of each amino acid a_j at each position i: emission probabilities $e_i(a_j)$
- Probability of a sequence segment x of length L given a model M: $P(x|M) = \prod_{i=1}^{L} e_i(x_i)$

```
Helix          DDDDDDDEEEEEEEEEEEEEEEEEEE                FFFFFFFFFFF
HBA_HUMAN  -DLS-----HGSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL-
HBB_HUMAN  GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL-
MYG_PHYCA  KHLKTEAEMKASEDLKKHGVTVLTALGAILKK----K-GHHEAELKPLAQSHATKH-
GLB3_CHITP AG-KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG-
GLB5_PETMA KGLTTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF-
LGB2_LUPLU LK-GTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-
GLB1_GLYDI SG----AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGN
Consensus    .  t    .. . v..Hg kv. a   a...l   d   . a l. l   H  .
```

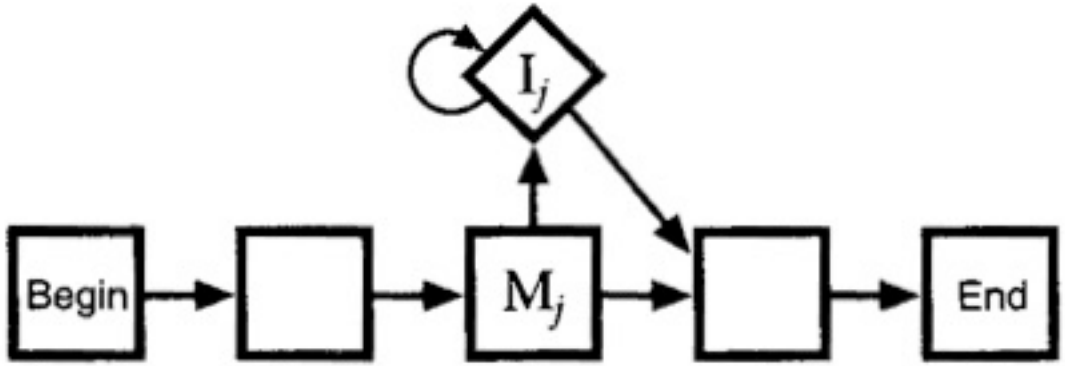
HMM archtechture: ungapped regions

$$\{e_{M_j}(a_i)\}$$

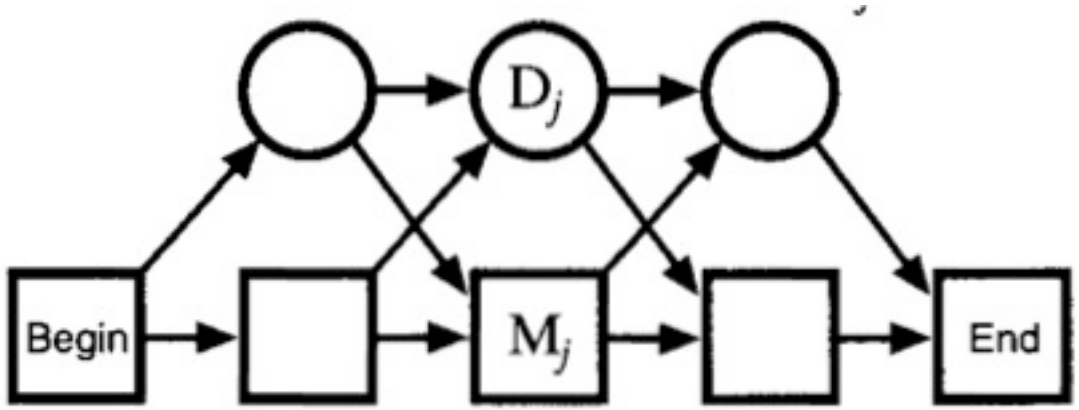
- **Alignment is trivial**
- Probability of a sequence given a model m: $P(x|m) = \prod_{i=1}^{L} e_i(x_i)$
- Probability of a sequence given a random model: $P(x) = \prod_{i=1}^{L} q_{x_i}$ (q_{x_i}: background frequency of amino acid x_i)
- Comparison to a random sequence: **log-odds** $S(x|m) = \prod_{i=1}^{L} \log \frac{e_i(x_i)}{q_{x_i}}$

Profile HMM: dealing with gaps

Insertions:

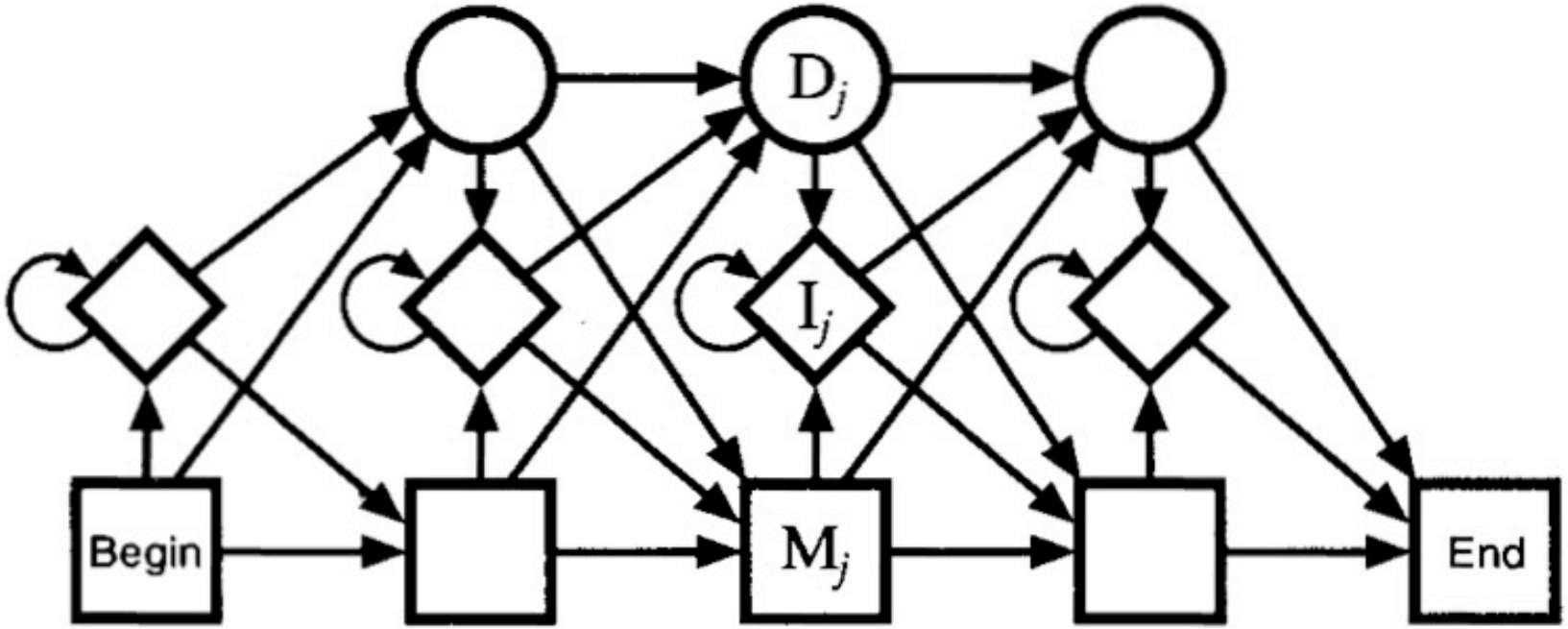


Deletions:

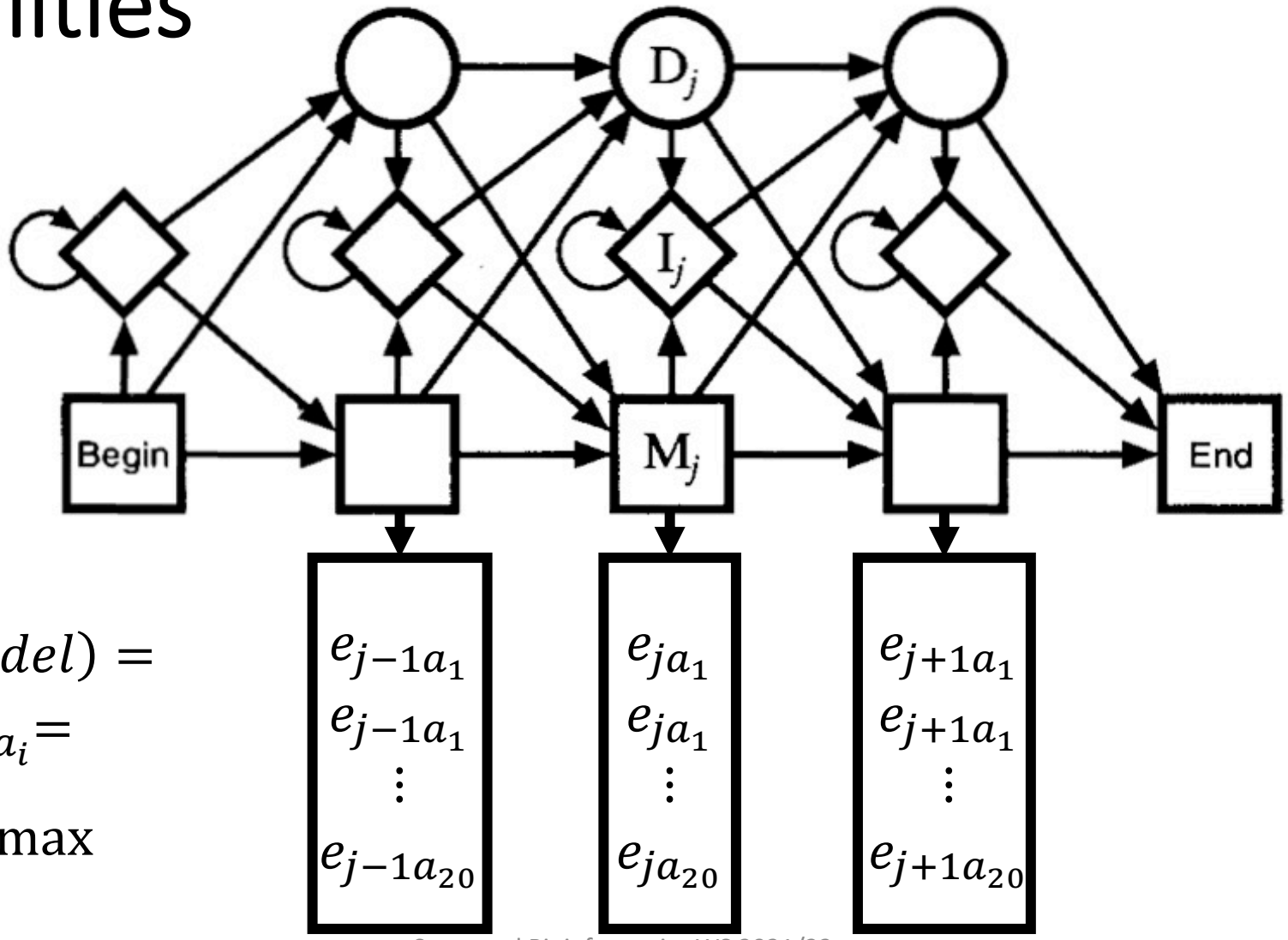


Profile HMM: full model architecture

- A separate HMM for every protein family
- # match states = alignment length

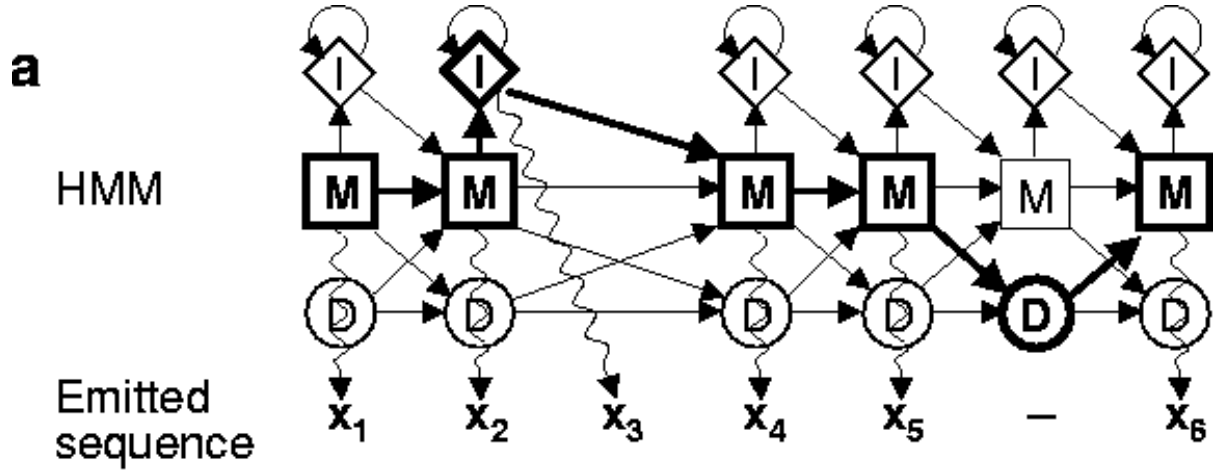


Profile HMM: full model with emission probabilities



$$P(sequence|model) =$$
$$= \sum P\left(X_j \middle| X_{j-1}\right) e_{ja_i} =$$
$$\sum a_{j,j-1} e_{ja_i} \rightarrow \max$$

Number of parameters



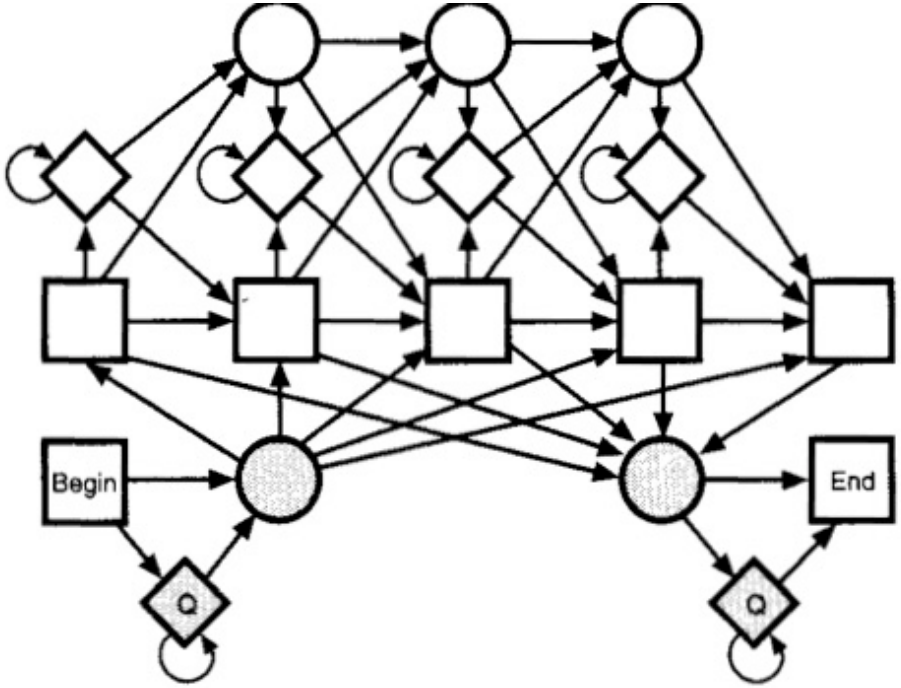
- Model length *L*
- Three hidden states for each position: **match**, **insertion**, **deletion** => 3 × *L* hidden states
- **match** and **insertion** states emit symbols (amino acid residues), each state may have its own distribution
- 4 × *L* transition probabilities
- 19 × 2 × *L* emission probabilities

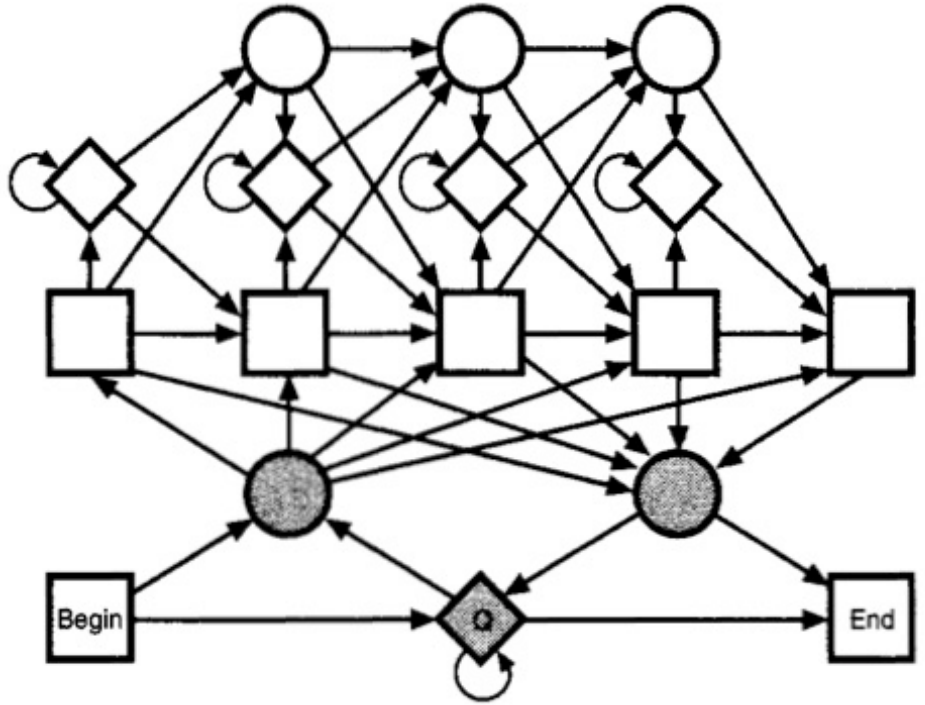
Probabilities can be estimated from the data

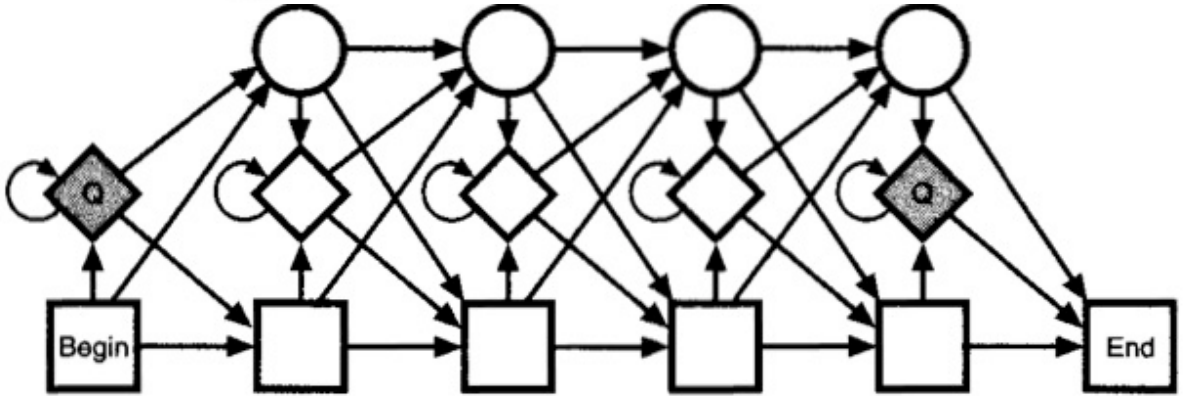
```
HBA_HUMAN   ...VGA--HAGEY...
HBB_HUMAN   ...V----NVDEV...
MYG_PHYCA   ...VEA--DVAGH...
GLB3_CHITP  ...VKG------D...
GLB5_PETMA  ...VYS--TYETS...
LGB2_LUPLU  ...FNA--NIPKH...
GLB1_GLYDI  ...IAGADNGAGV...
              ***  *****
```

- Suppose we have a large enough alignment
  - Consensus sequence length: how many ungapped or reasonably ungapped positions do we have — heuristic
  - Transition and emission probabilities:
    - $\boldsymbol{a_{kl}} = \frac{\boldsymbol{A_{kl}}}{\sum_{\boldsymbol{l'}} \boldsymbol{A_{kl'}}}$, A_{kl}: number of observed transitions, k, l: all states
    - $\boldsymbol{e_k(a)} = \frac{\boldsymbol{E_k(a)}}{\sum_{\boldsymbol{a'}} \boldsymbol{E_k(a')}}$, $E_k(a)$: number of occurrences of a in state k
  - Can be updated as we gather more proteins
  - Can be mixed with background frequencies: **pseudocounts**

HMMs for local alignments







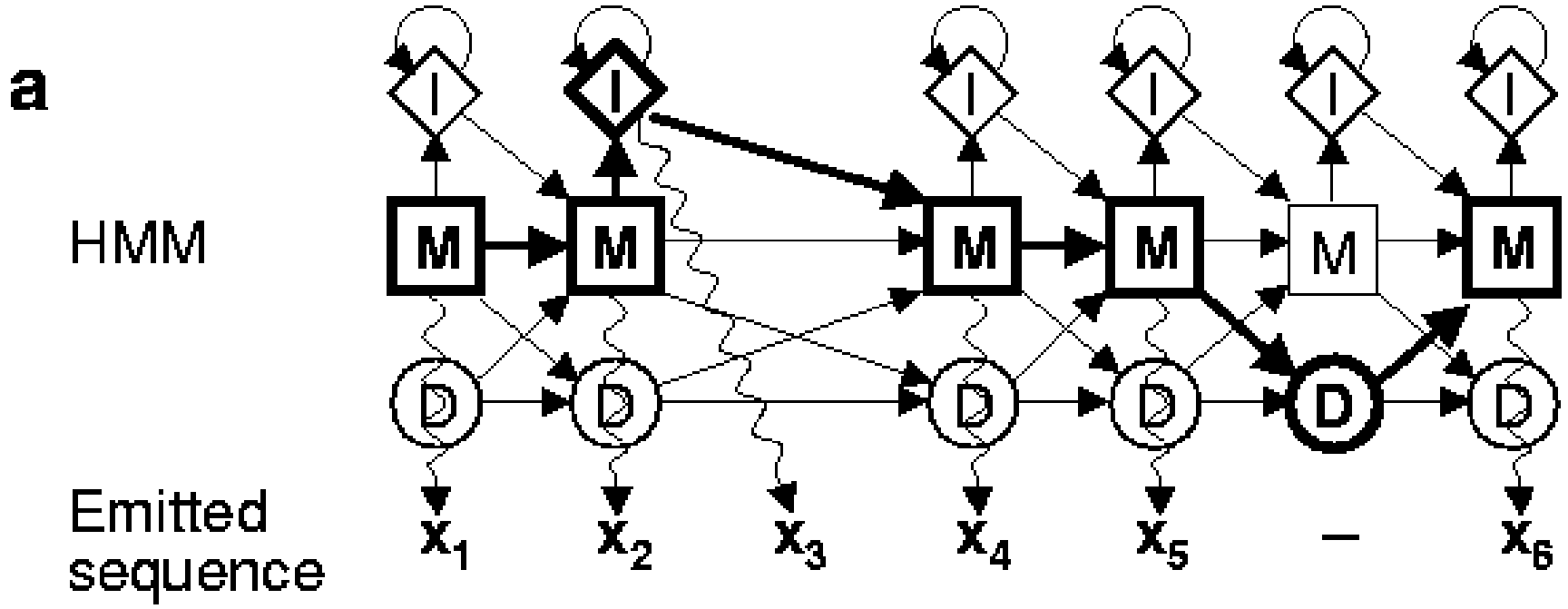
Searching with profile HMMs

- Given a model *M*, calculate
  - probability of a particular alignment π^* of a sequence *x* $P(x, \pi^* | M)$
  - full probability of *x* summed over all possible alignments $P(x|M)$
- Given a set of models $\{M_1, \ldots, M_K\}$ and a database of sequences $\{x_1, \ldots, x_N\}$, group them to models
- Compare to the score of the model of a random sequence $P(x|R) = \prod_i q_{x_i}$

Profile-profile sequence search

HMM-sequence alignment

- Find the best path through a specific model



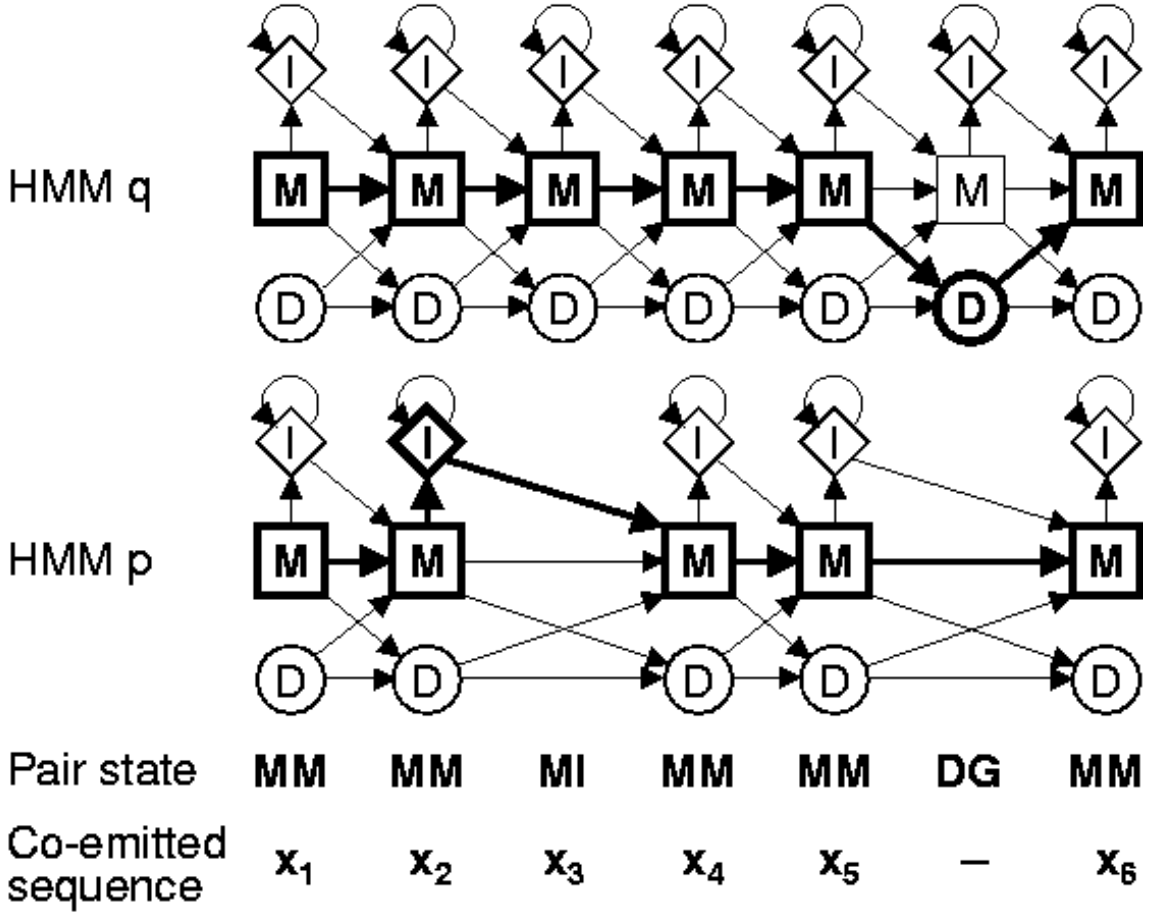
HMM-HMM profile alignment

- **Align two HMMs** (built for two groups of sequences)
- A path through one model is matched to a path through another model
- Additional states: **gaps** in the whole model
  - exactly like alignment deletions
- States can be matched only in a specific manner:
  - no matches between **insertion** an **deletion** states, **insertion** and **gap** states
- Transitions between insertion and deletion states are improbable and do not influence scoring much

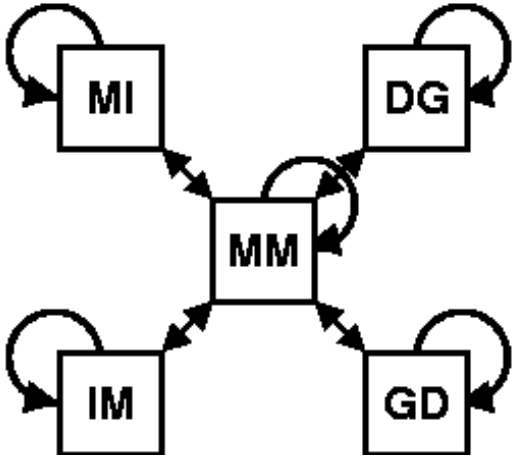
Söding, 2005

HMM-HMM profile alignment

- **Söding, Bioinformatics, 2005: "Protein homology detection by HMM–HMM comparison"**



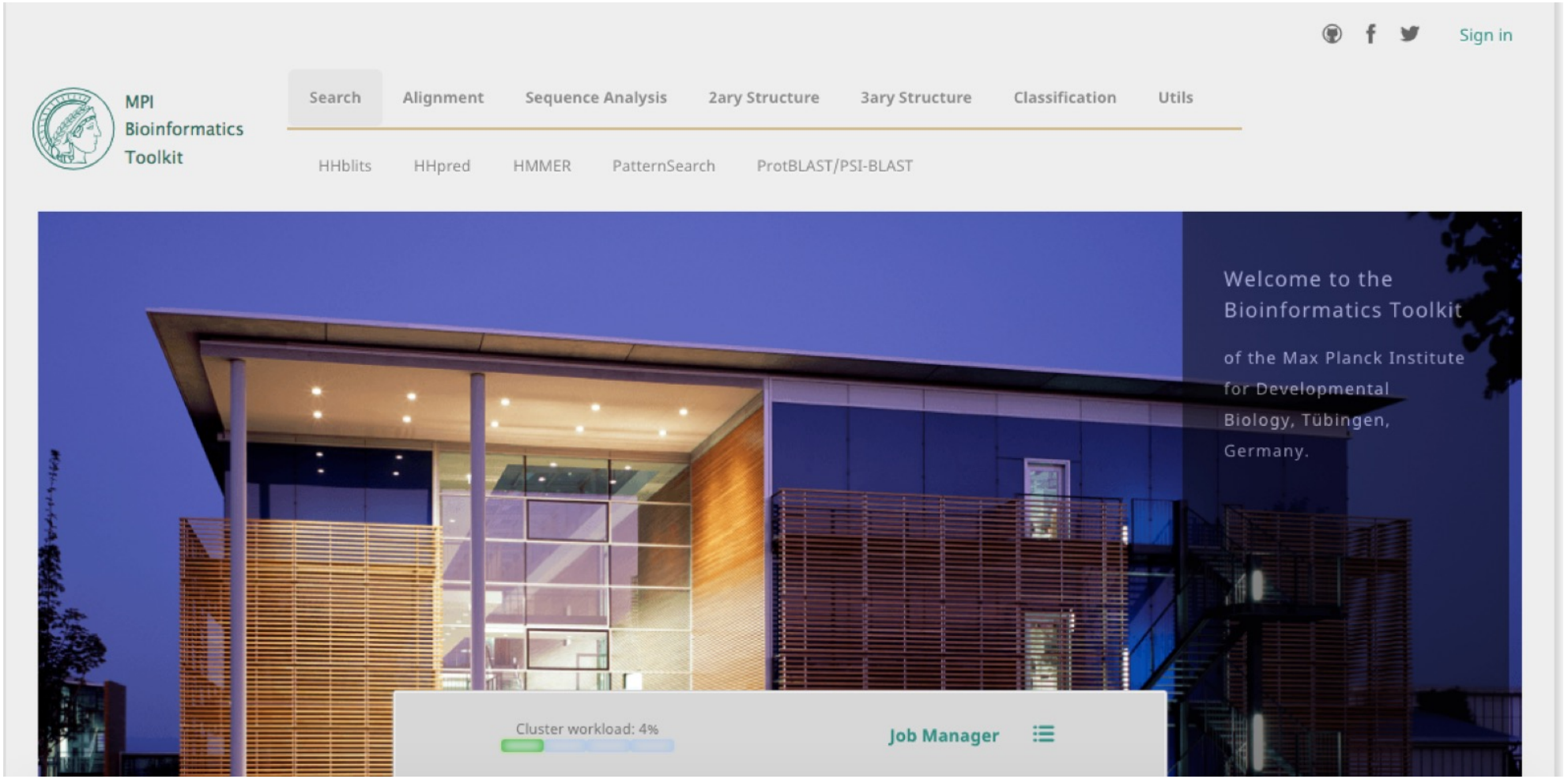
Allowed pair state transitions



Implementations: HHsearch (Söding, 2005) / HHblits (Remmert et al., 2011)

- Implement HMM-HMM profile sequence search
- Can use single sequence or an alignment as input
- Comes with a pre-calculated collection of HMMs
- Available for download or online as a part of the Tübingen Bioinformatics Toolkit at https://toolkit.tuebingen.mpg.de/

Implementations: HHsearch (Söding, 2005) / HHblits (Remmert et al., 2011)



Protein structure similarity and evolution

Root Mean Square Deviation (RMSD)

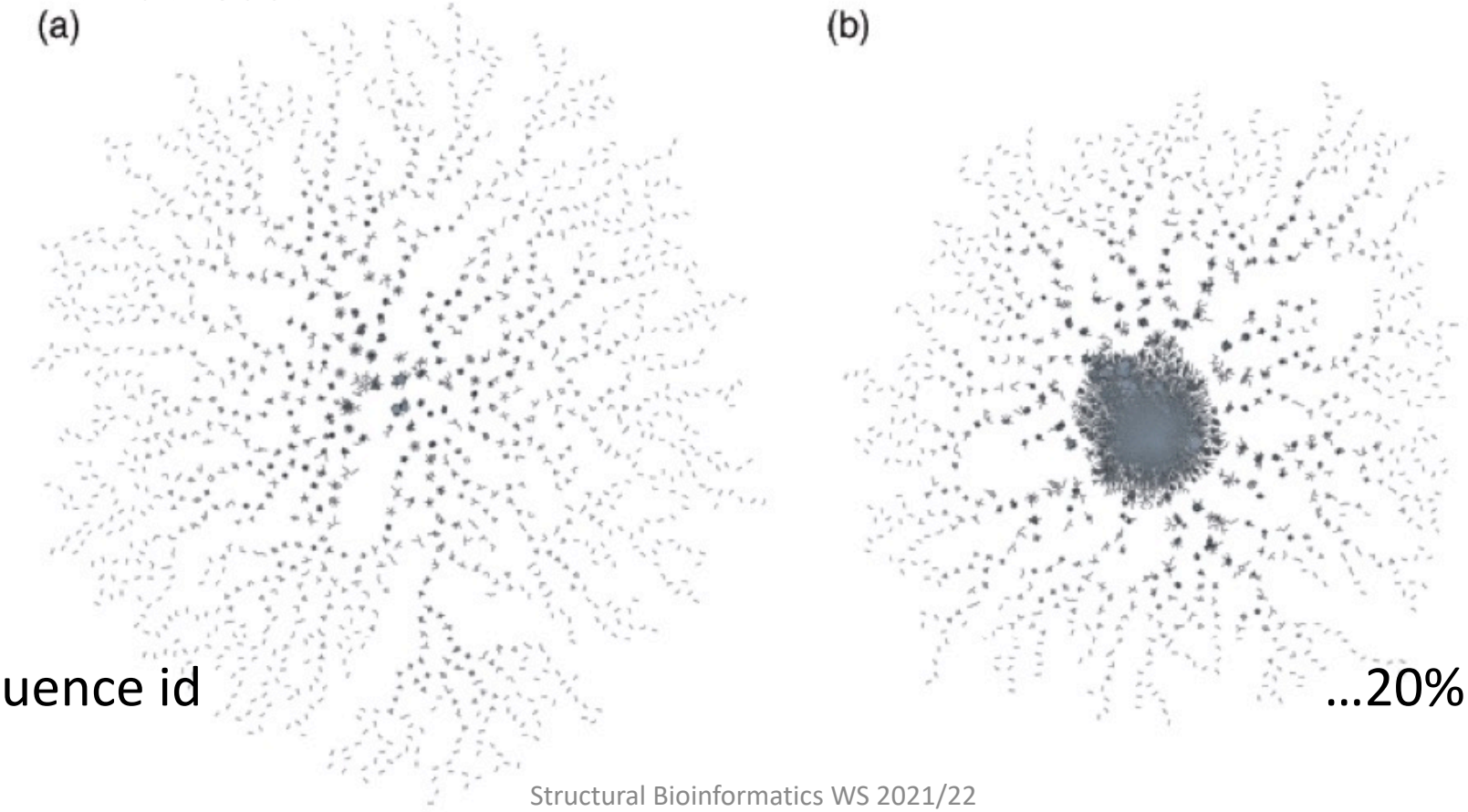
- The simplest measure of **similarity for 3D structures** (more: lecture 8)

- $$RMSD = \sqrt[2]{\frac{\sum_{i=1}^{N}\left(\left(x_i^1 - x_i^2\right)^2 + \left(y_i^1 - y_i^2\right)^2 + \left(z_i^1 - z_i^2\right)^2\right)}{N}},$$

  where (x_i^1, y_i^1, z_i^1) are coordinated of the *i*-th atom of structure 1 and (x_i^2, y_i^2, z_i^2) are coordinated of the *i*-th atom of structure 2

Protein fold space

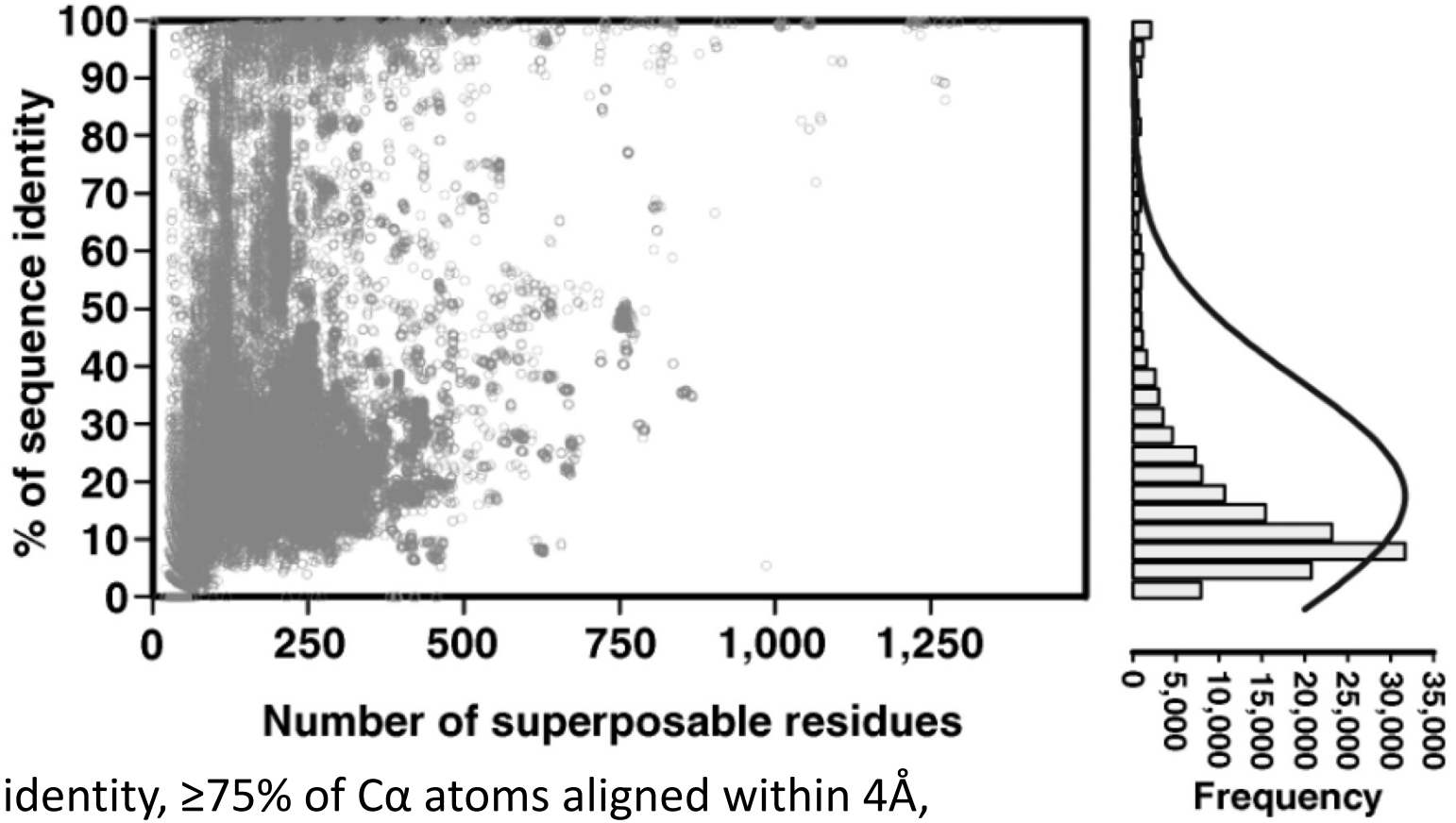
Vertex: protein; edge: >40% Cα can be superimposed at 4Å and result in an alignment with at least…



…40% sequence id

…20% sequence id

Sequence vs. structural similarity (as of 2007)

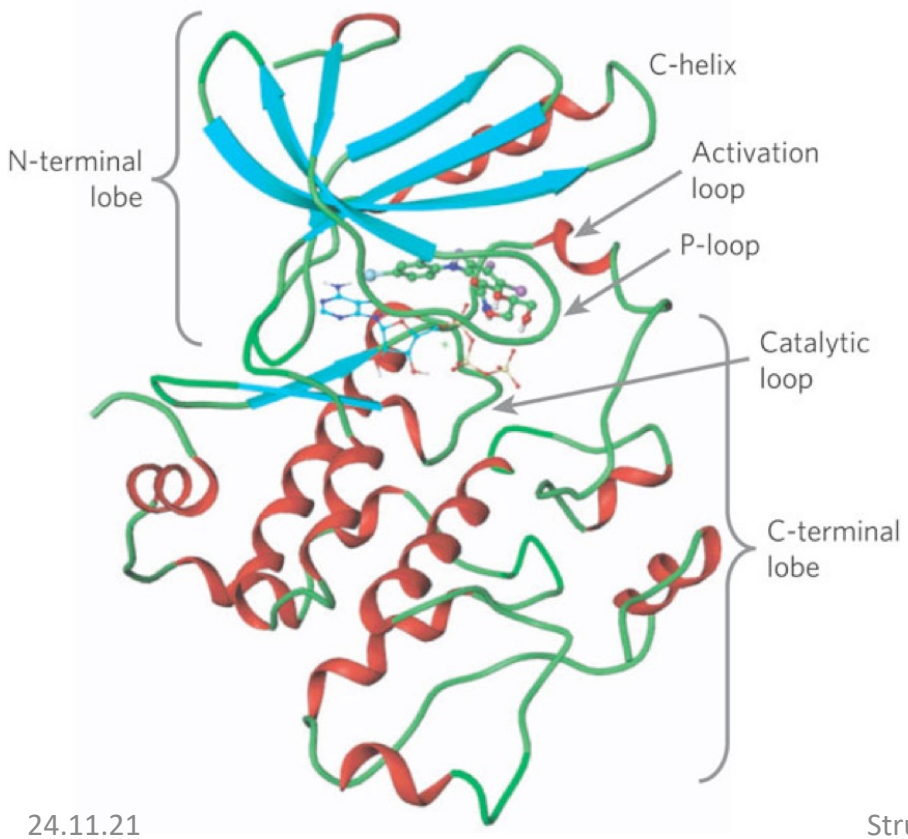


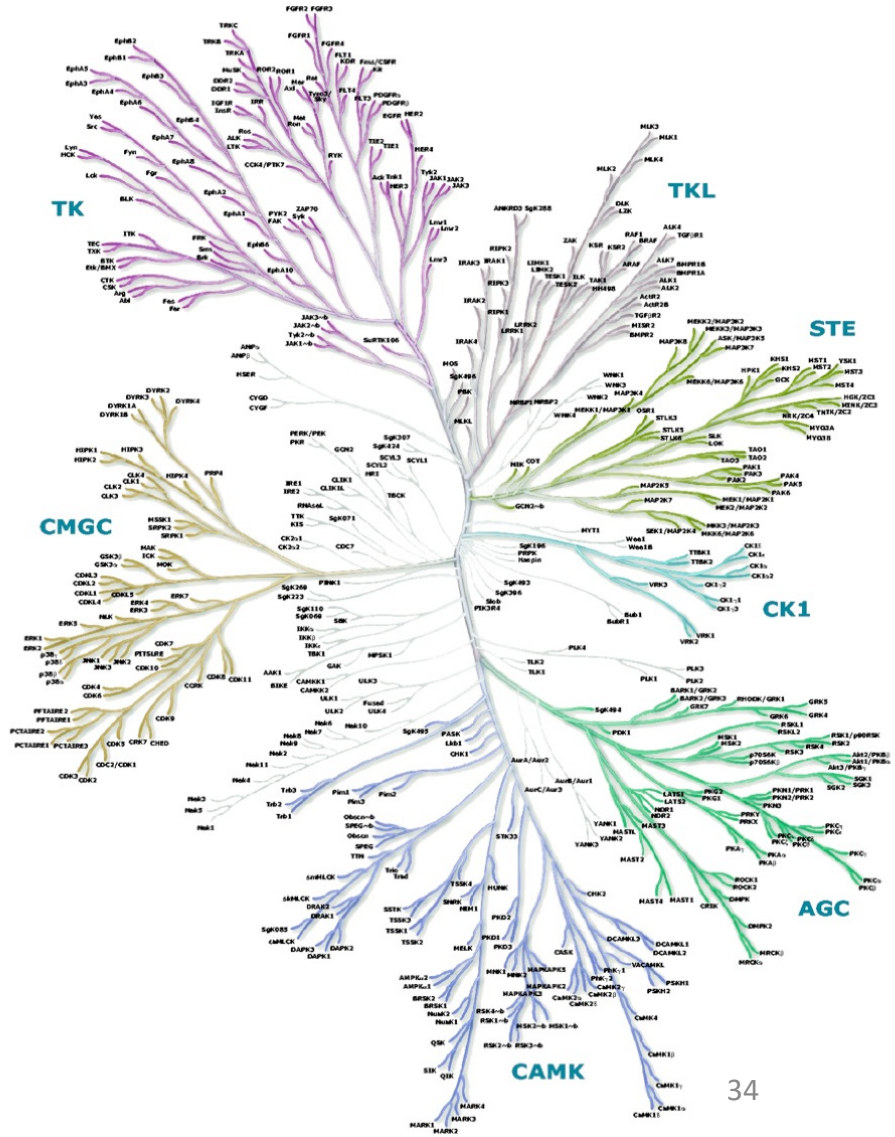
≥20% sequence identity, ≥75% of Cα atoms aligned within 4Å,
≤3Å Cα RMSD, and ≤50 residues difference in length:
159,777 pair-wise structural alignments

"Structural bioinformatics", ed Bourne and Gu

Large protein families, e.g. the human kinome

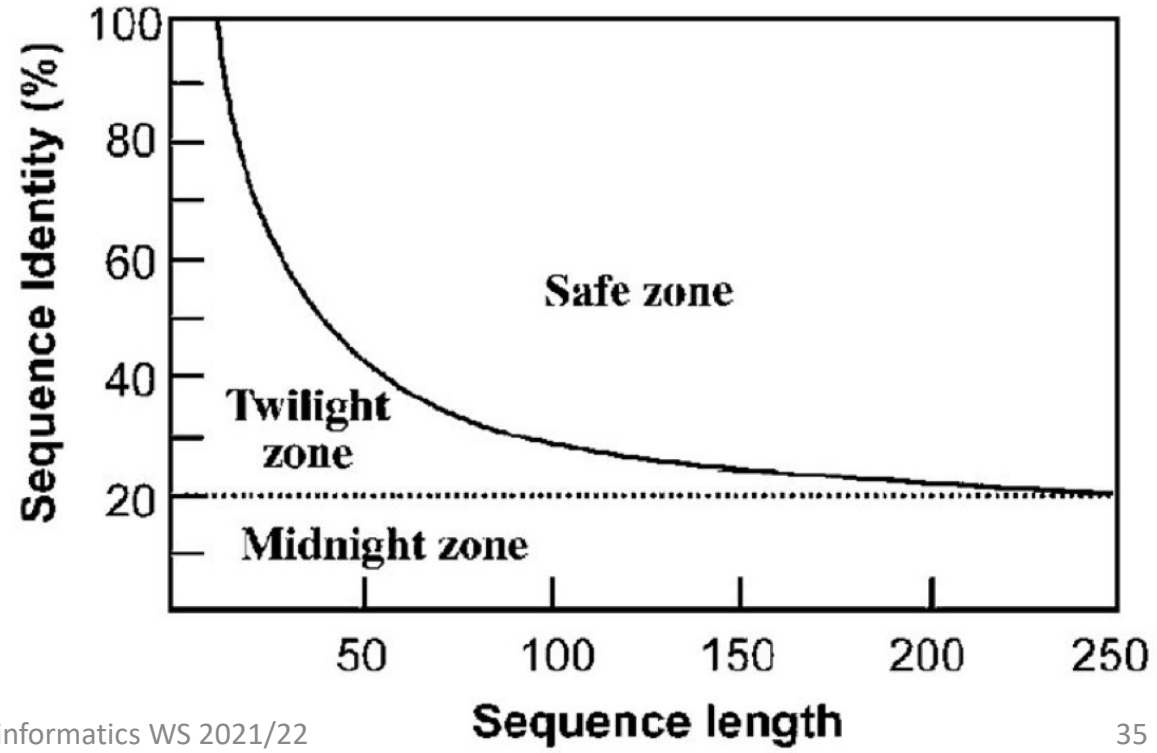
Some kinases have as little as 18% identity, yet same kinase fold





Protein similarity and evolutionary relatedness

- **Similar sequences** => evolutionary related
- **Similar structures** => ???
- **But**: profile sequence similarity search get more and more sensitive (also as the databases grow)

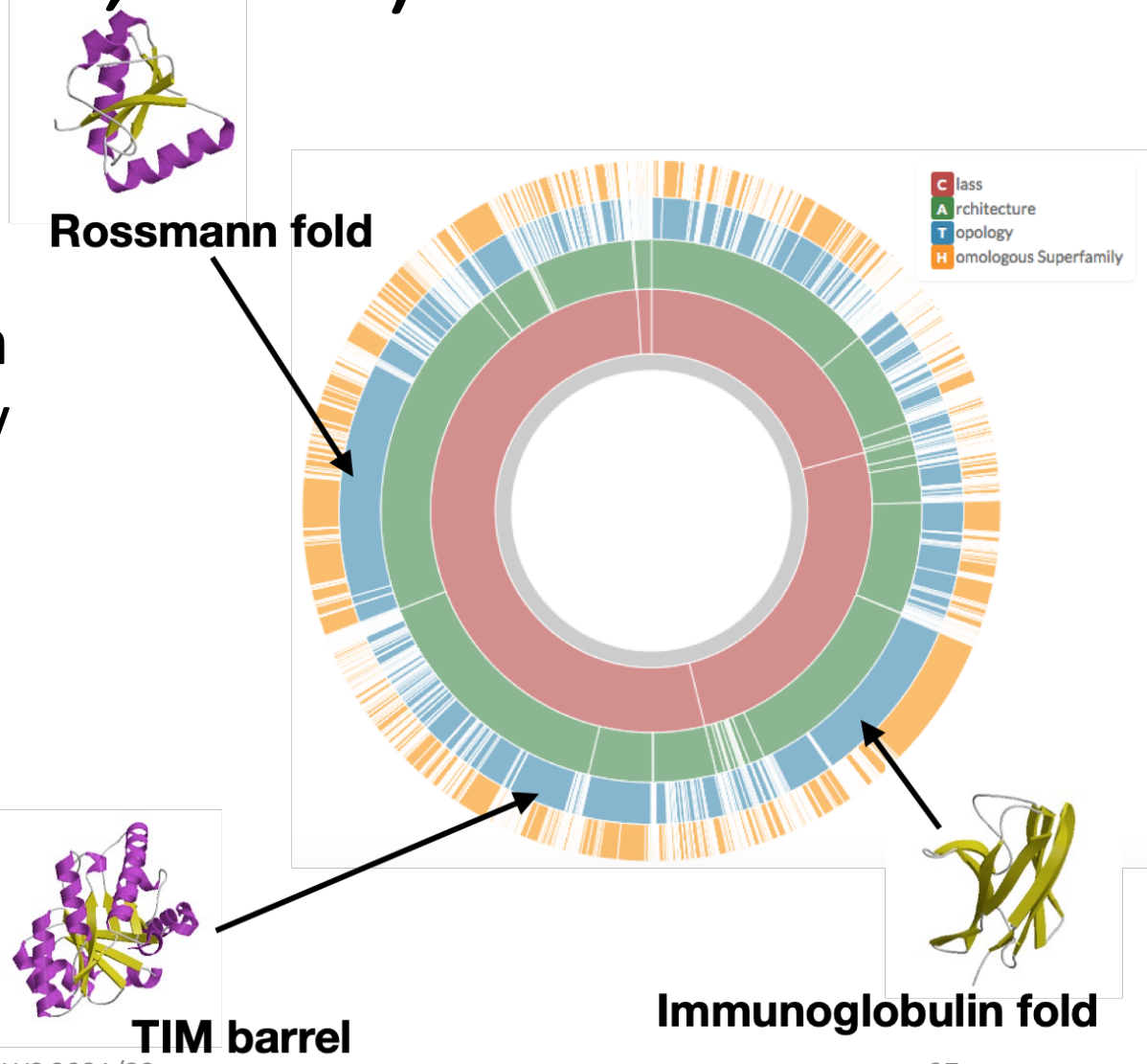


Protein evolution —
The basic building blocks of life

- Andrei N. Lupas
- Vikram Alva
  - Max Planck Institute for Developmental Biology
- Johannes Söding
  - Max Planck Institute for Biophysical Chemistry

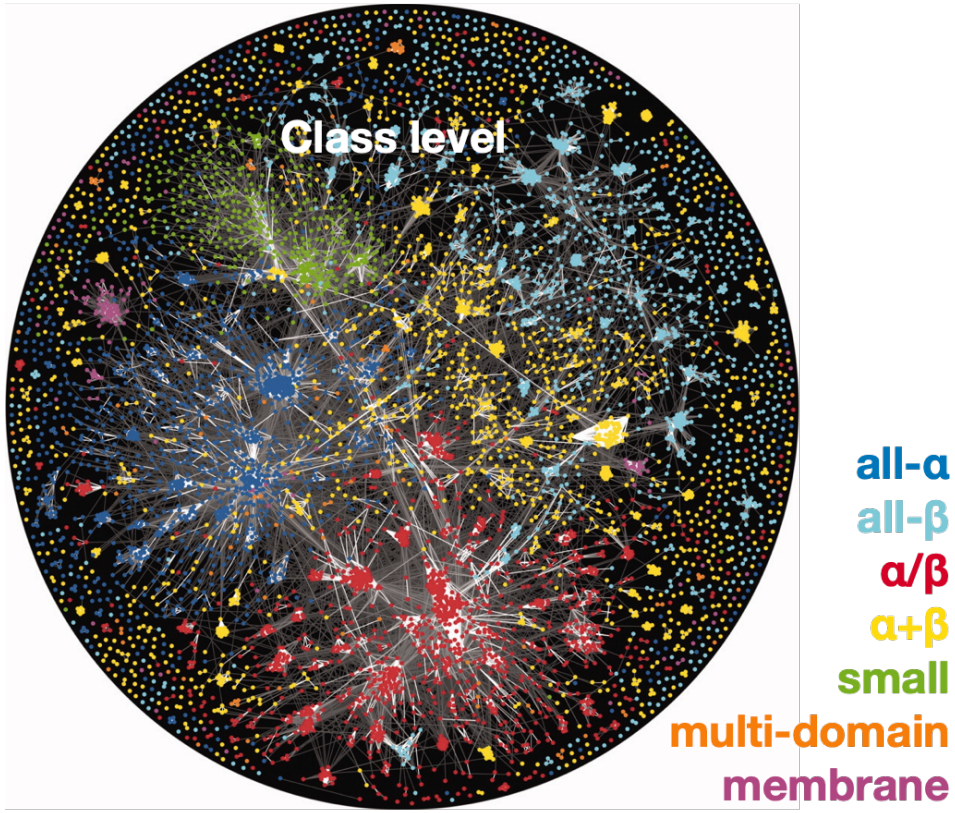
A galaxy of folds (Alva et al., 2010)

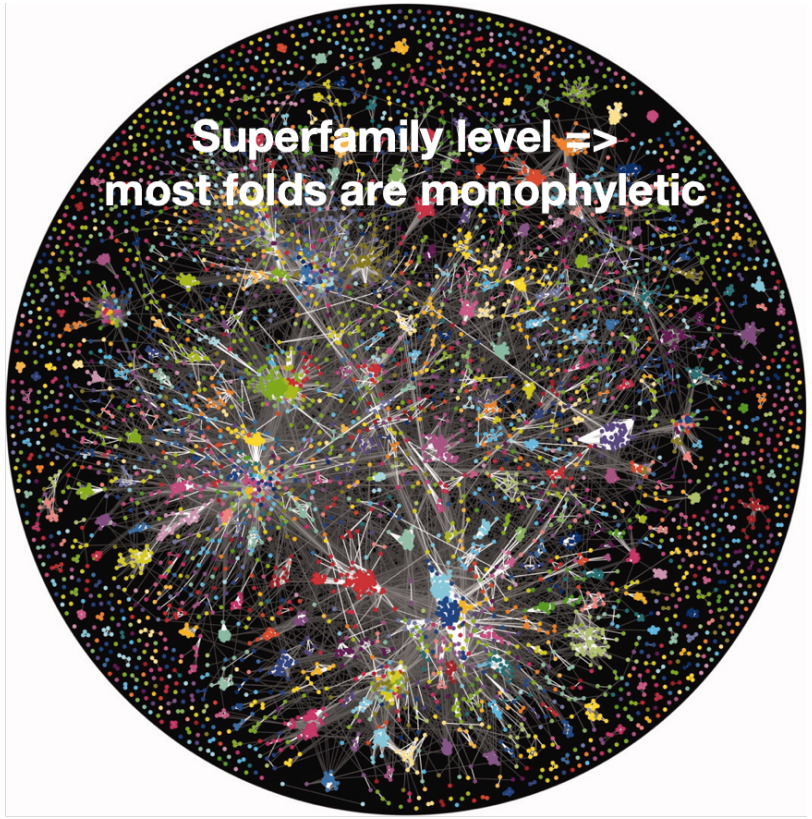
- ~1,000 protein folds, ~10,000 protein families in nature, some of them very populated
- Did they evolve independently?
- Are structural similarities due to chance or common ancestry?



A galaxy of folds (Alva et al., 2010)

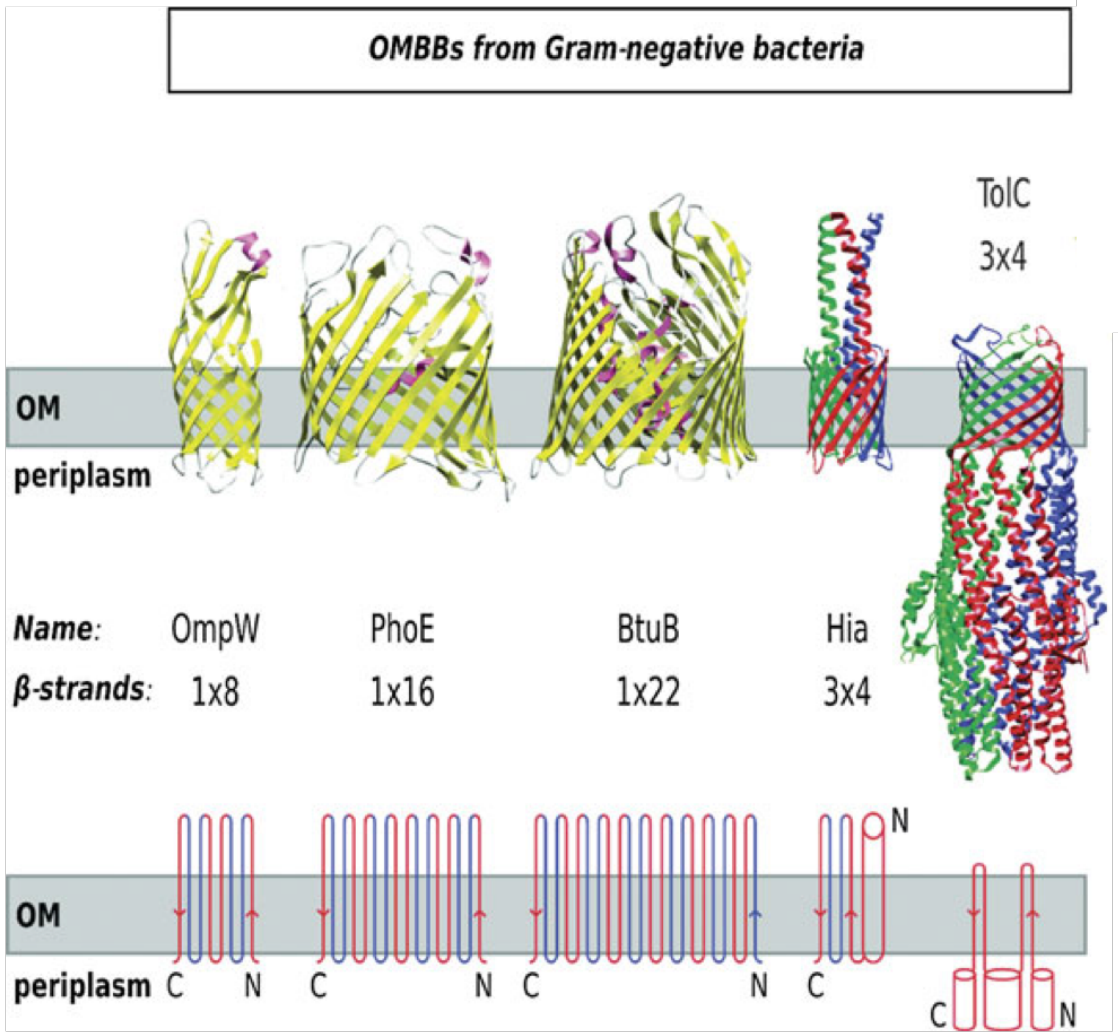
- Filter SCOP down to 20% sequence identity (midnight zone)
- HHsearch between different SCOP families





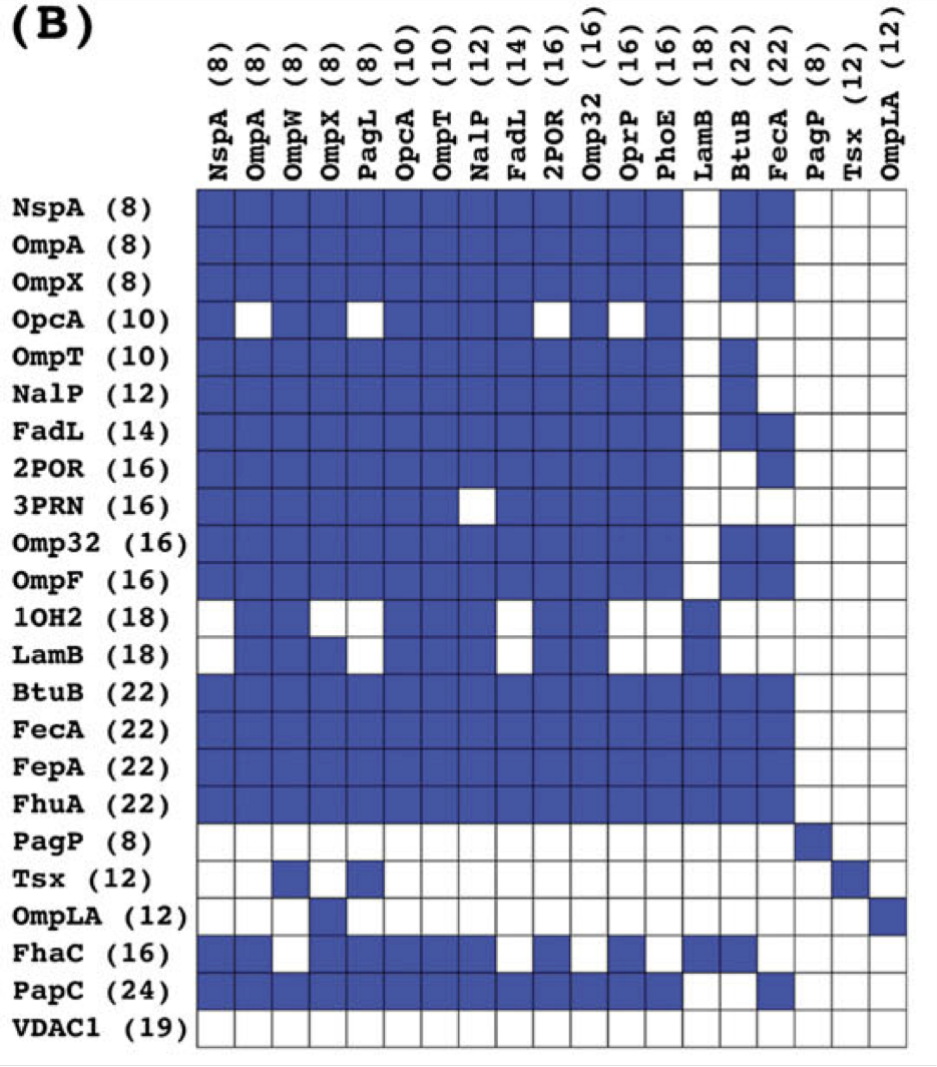
Outer membrane proteins (Remmert et al., 2010)

- Outer membrane β-barrels
  - channels in the outer membrane of Gram-negative bacteria
- Link different groups of known related OMBBs
  - profile search (HHsearch)
  - cluster hits
  - internal repeat detection



Outer membrane proteins (Remmert et al., 2010)

- Most OMBB groups find each other but not other proteins
- Similarity via ββ-hairpin
- Found hits are structurally similar
- Sequence similarity not due to structural convergence
- Many proteins have a self-repeat pattern

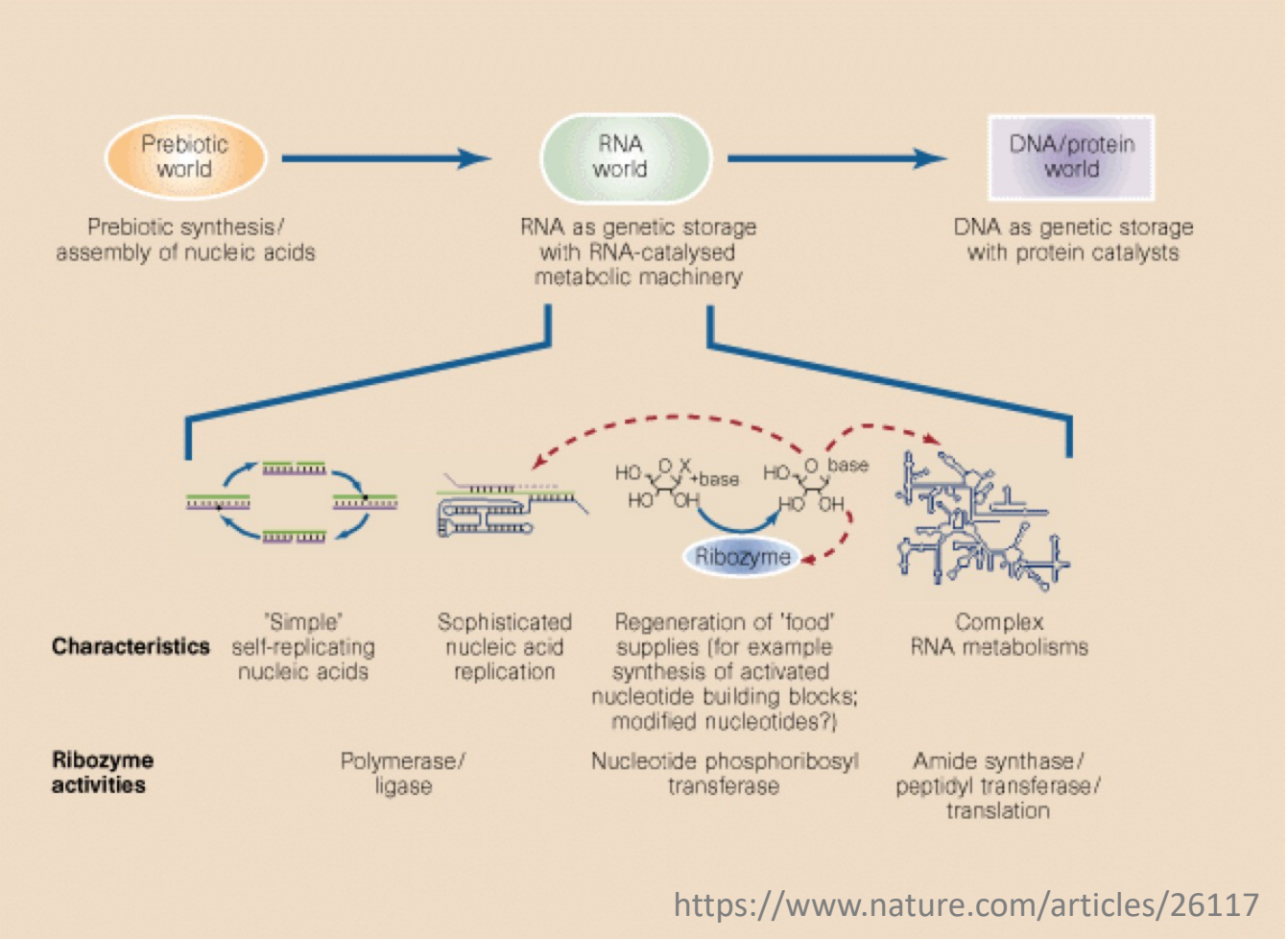


A vocabulary of ancient peptides (Alva et al., 2015)

- **Domains**: independently folding 3D arrangements of secondary structure elements
- Domains are units of protein evolution
- How did domains evolve?

The origin of life: the RNA world hypothesis

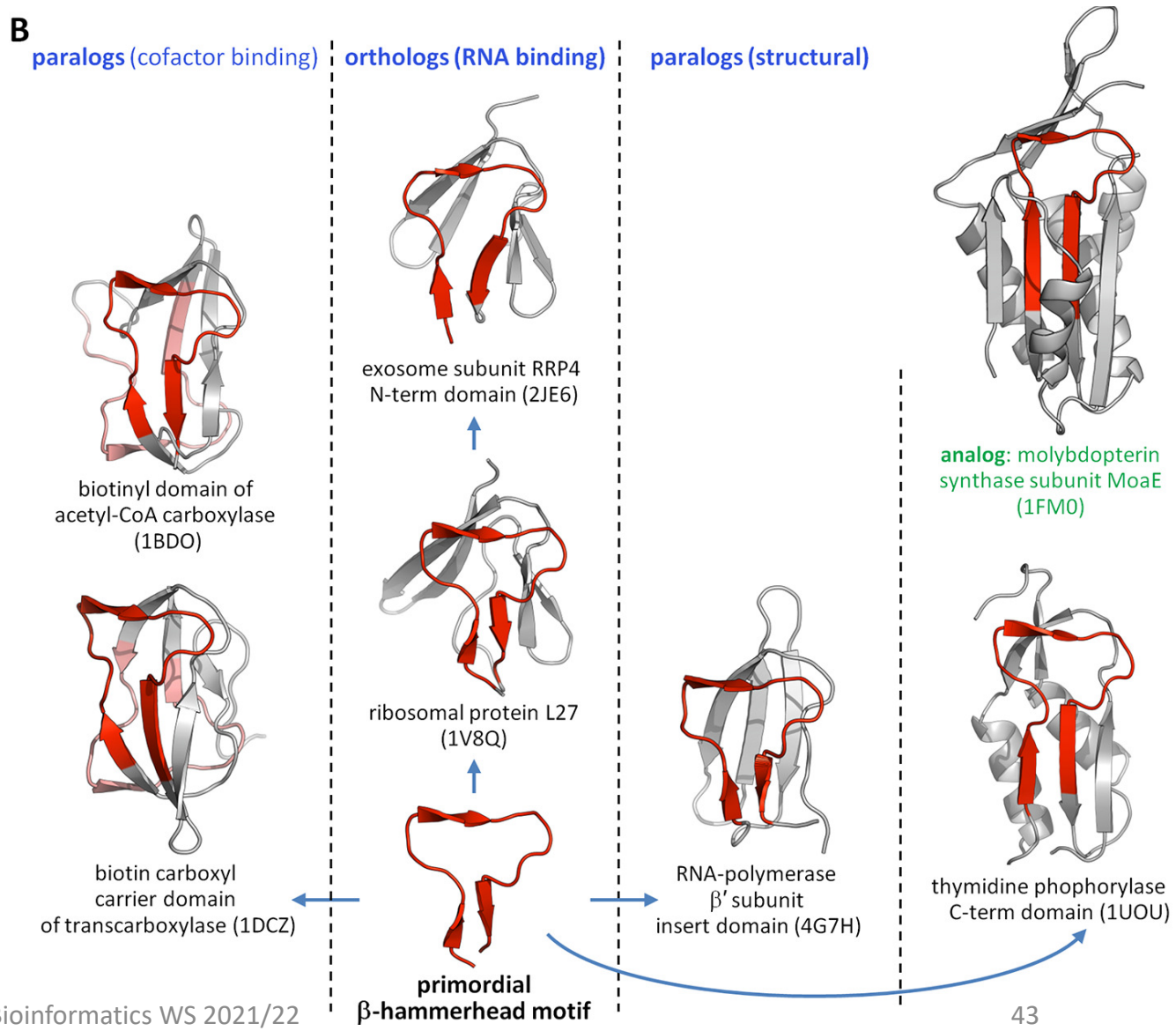
- An intermediate stage in the origin of life
- Self-replicating RNA molecules as precursors of modern replicons (consisting of DNA, RNA, proteins)
- Catalytic capacity of RNAs is limited, RNAs are not very stable => recruit peptides to expand their functions



https://www.nature.com/articles/26117

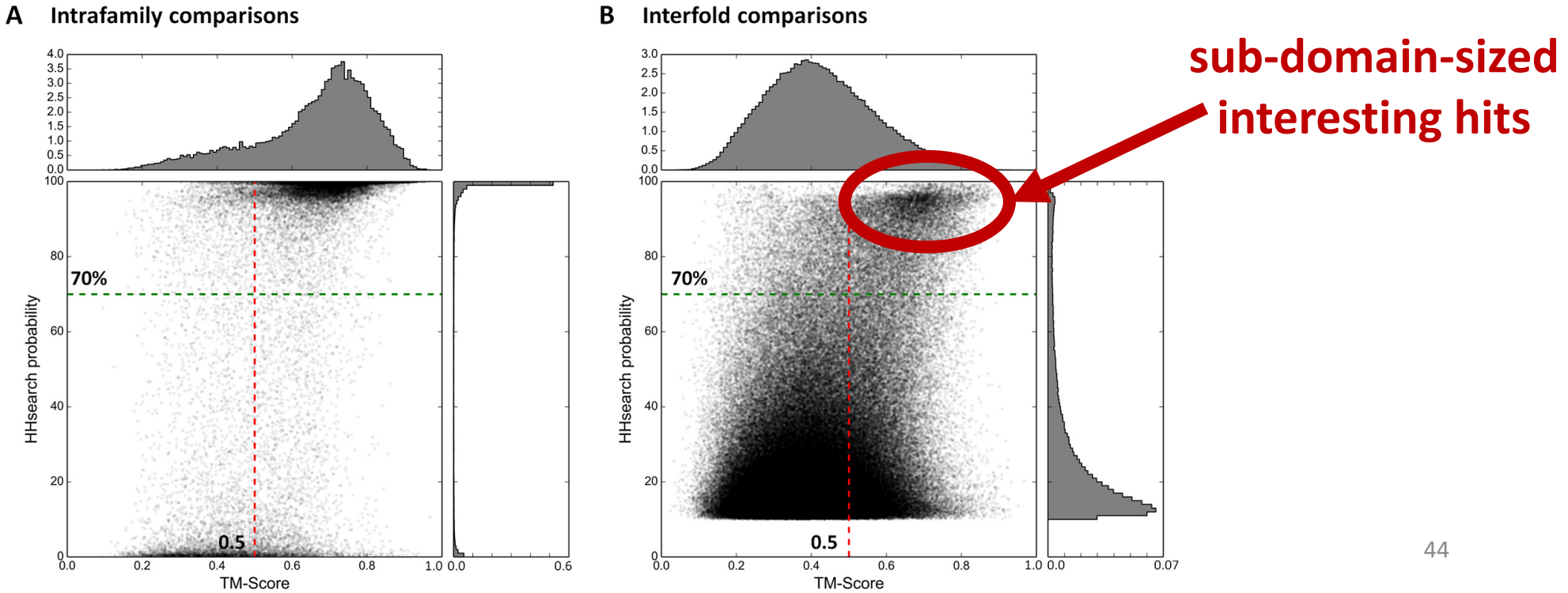
A vocabulary of ancient peptides (Alva et al., 2015)

- **Aim**: discover the ancestors of these ancient peptides, families, functions
- **Model**:



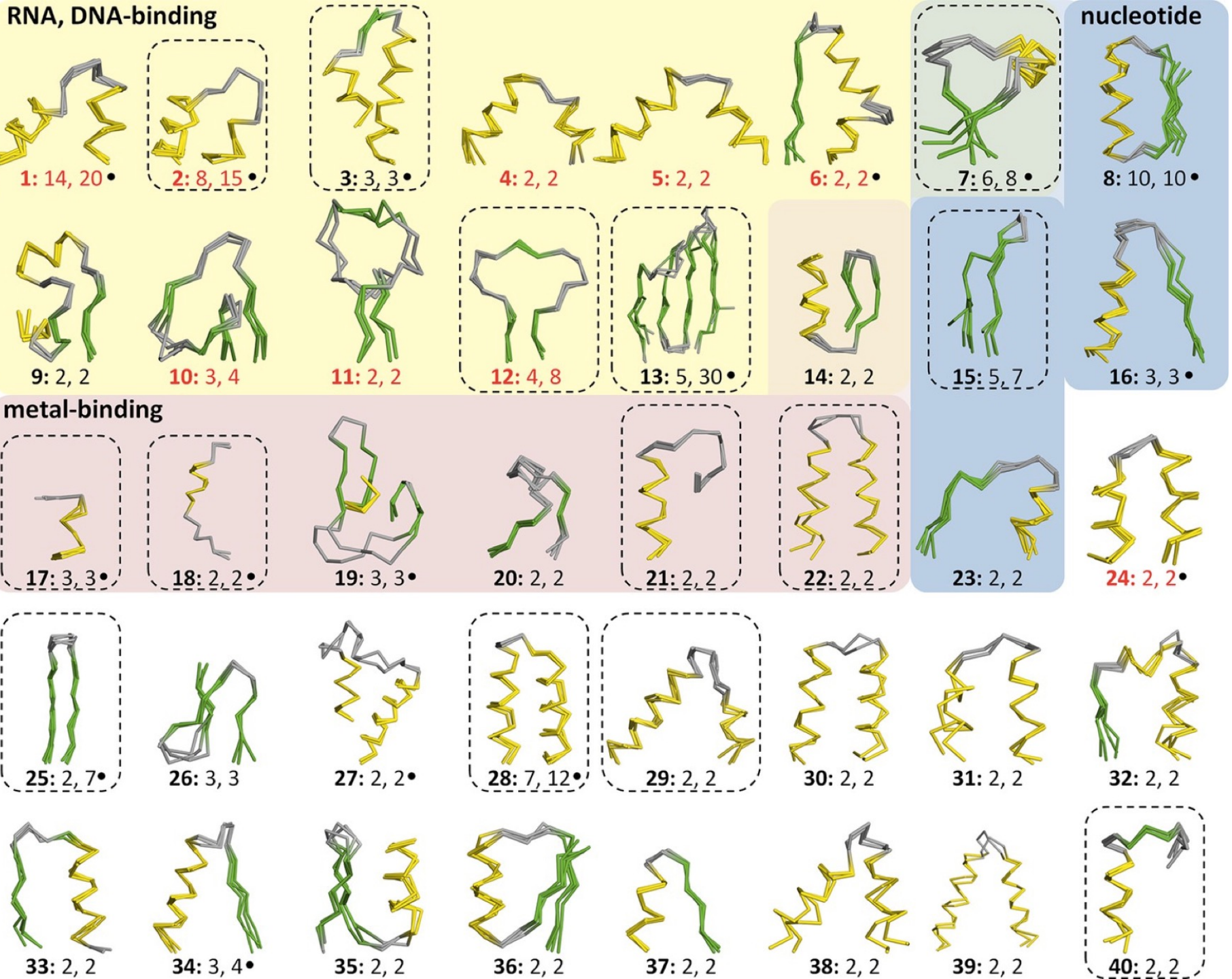
A vocabulary of ancient peptides (Alva et al., 2015)

- **Approach**:
  - filter SCOPe down to 30% sequence identity, HHsearch
  - clustering of 3D structures



A vocabulary of ancient peptides (Alva et al., 2015)

- **Result**: 65 structural clusters
  - 25 from related folds classified separately (e.g. due to permutation) or errors in definition of fold boundaries
  - => 40 primordial peptides



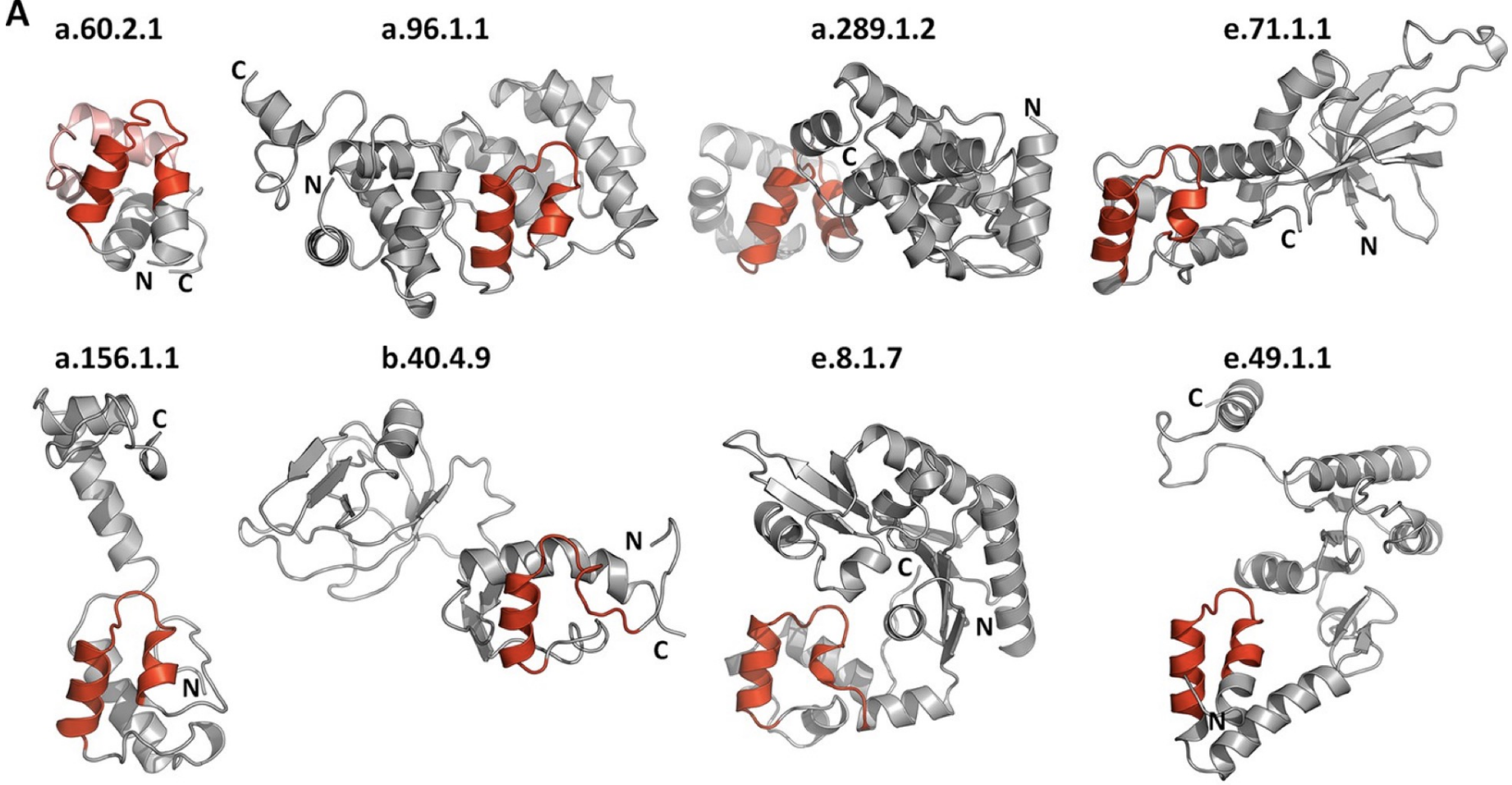
Legend:

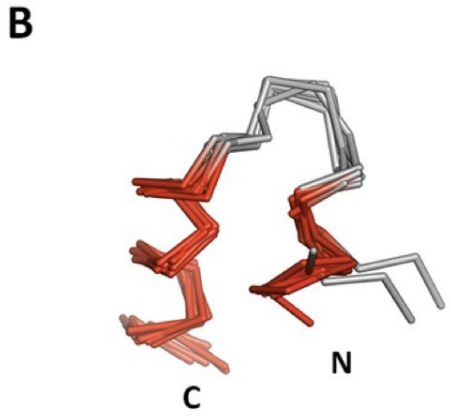
#: folds, superfamilies

•: reported before

found in the ribosome

form fold by repetition



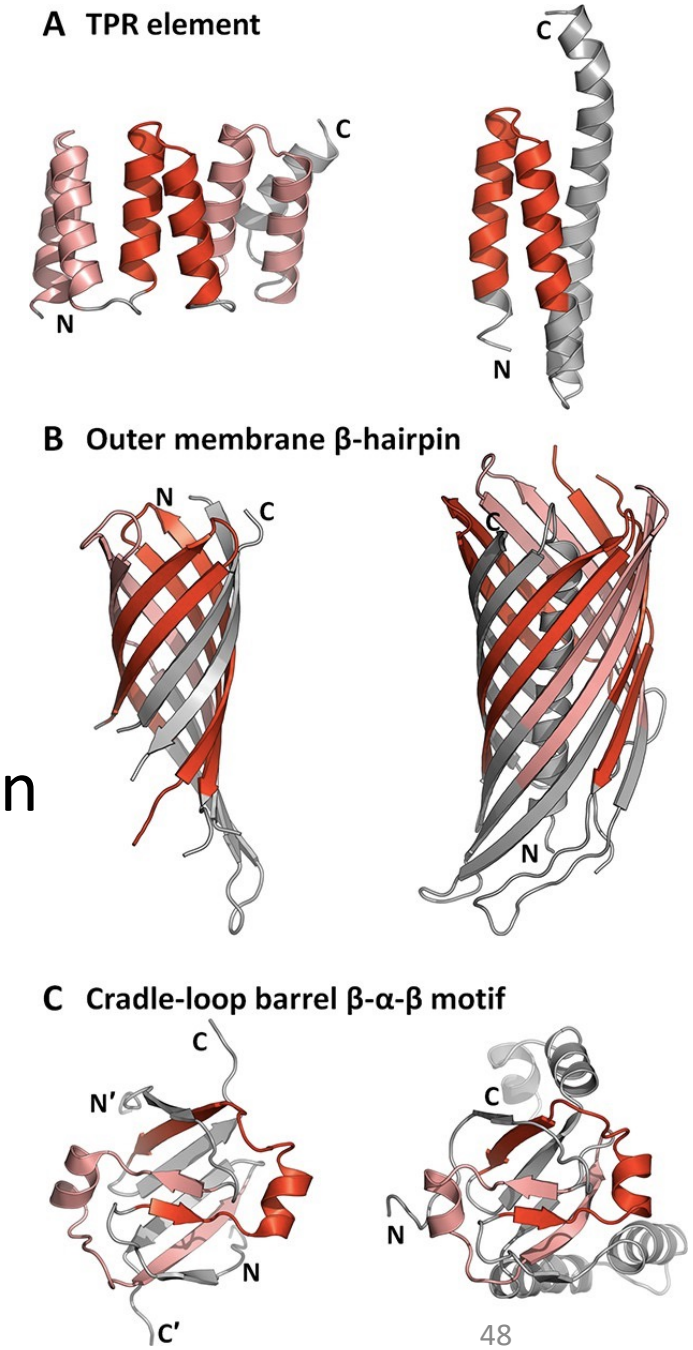


C

```
                                  hhhhhccccchhhhhhhhhh
1IXR (A:71-92)      a.60.2.1      FELLLSVSGVGPKVALALLSAL
1IXR (A:106-127)    a.60.2.1      ARLLTSASGVGRRLAERIALEL
1ORN (A:109-130)    a.96.1.1      RDELMKLPGVGRKTANVVVSVA
2UUB (M:16-37)      a.156.1.1     DVALTYIYGIGKARAKEALEKT
2P6R (A:631-652)    a.289.1.2     LLELVRIRHIGRVRARKLYNAG
1GM5 (A:114-135)    b.40.4.9      STDIQYAKGVGPNRKKKLKKLG
1JX4 (A:177-198)    e.8.1.7       ELDIADVPGIGNITAEKLKKLG
3VDP (A:9-30)       e.49.1.1      IEELSKLPGIGPKTAQRLAFFI
2I5H (A:129-150)    e.71.1.1      MHQLELLPGVGKKMMWAIIEER
```

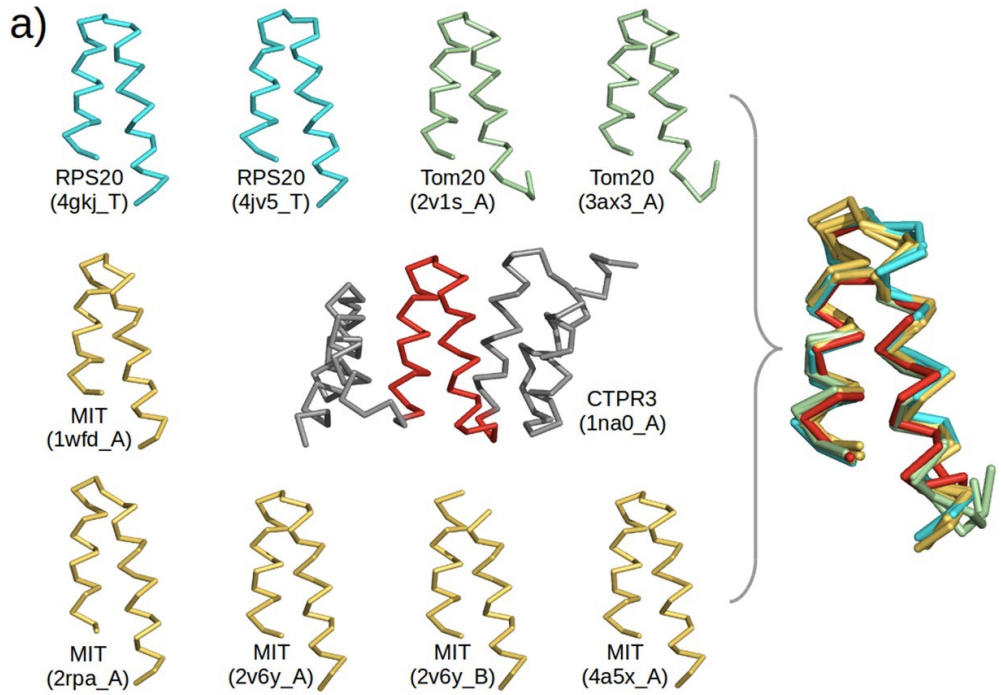
A vocabulary of ancient peptides (Alva et al., 2015)

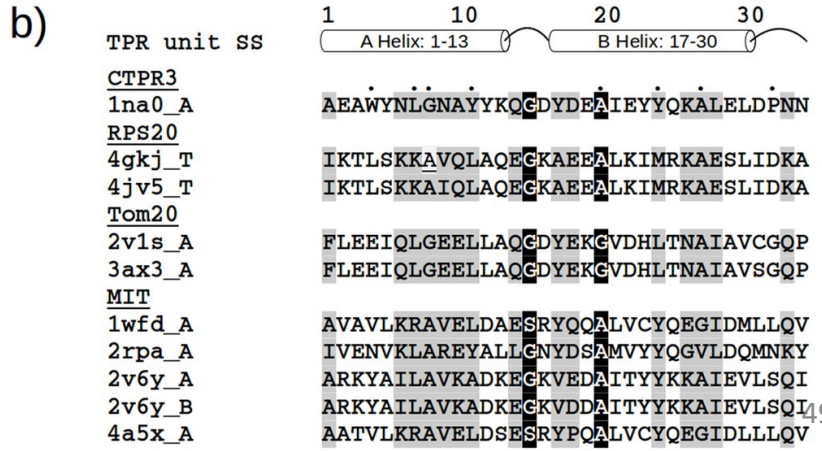
- **Conclusions**:
  - a lot of fragments found in proteins to contact nucleic acids, in agreement with the RNA world hypothesis
  - repetition is the driving force for the fold formation
  - identified fragments often constitute the conserved folding core of proteins; less conserved fragments can be attached



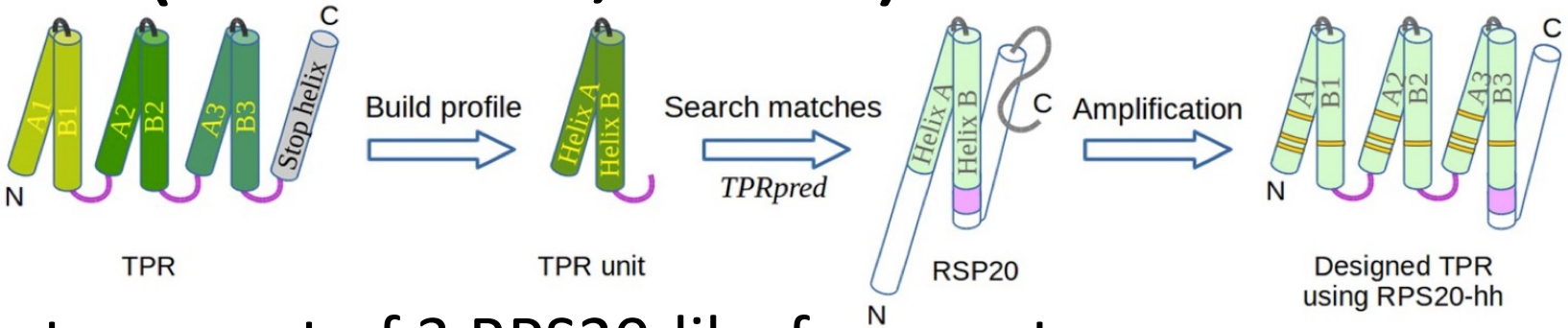
Transition from disorder to fold: TPR unit (Zhu et al., 2016)

- A conserved (HHsearch-level) fragment
  - found in proteins that assume stable 3D structure (CTRP3, MIT, etc.)
  - found in RPS20: ribosomal protein that gets folded only upon binding to the ribosome

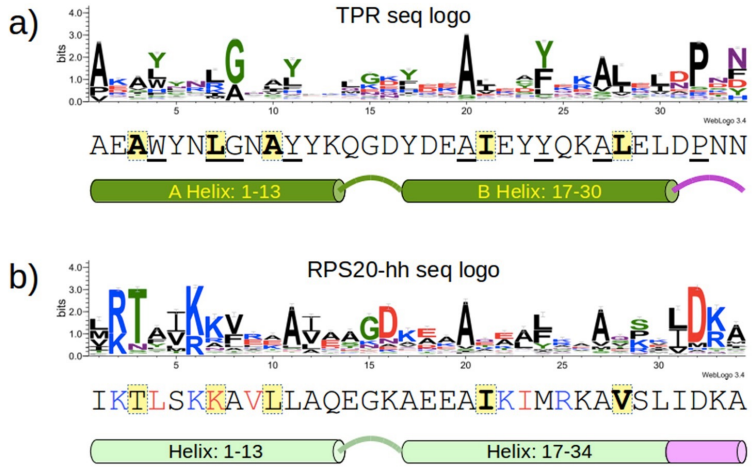


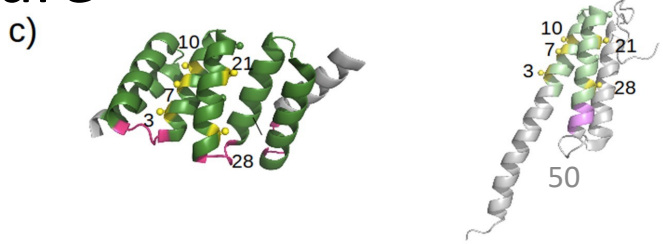


Transition from disorder to fold: TPR unit (Zhu et al., 2016)



- Construct a repeat of 3 RPS20-like fragments
- It is disordered in solution
- Introduce up to 5 mutations => 4 of 5 designed proteins fold well
  - due to improved contacts between the repeats
- Confirmed by an experimentally resolved 3D structure
- The mutations work in the original RPS20 context



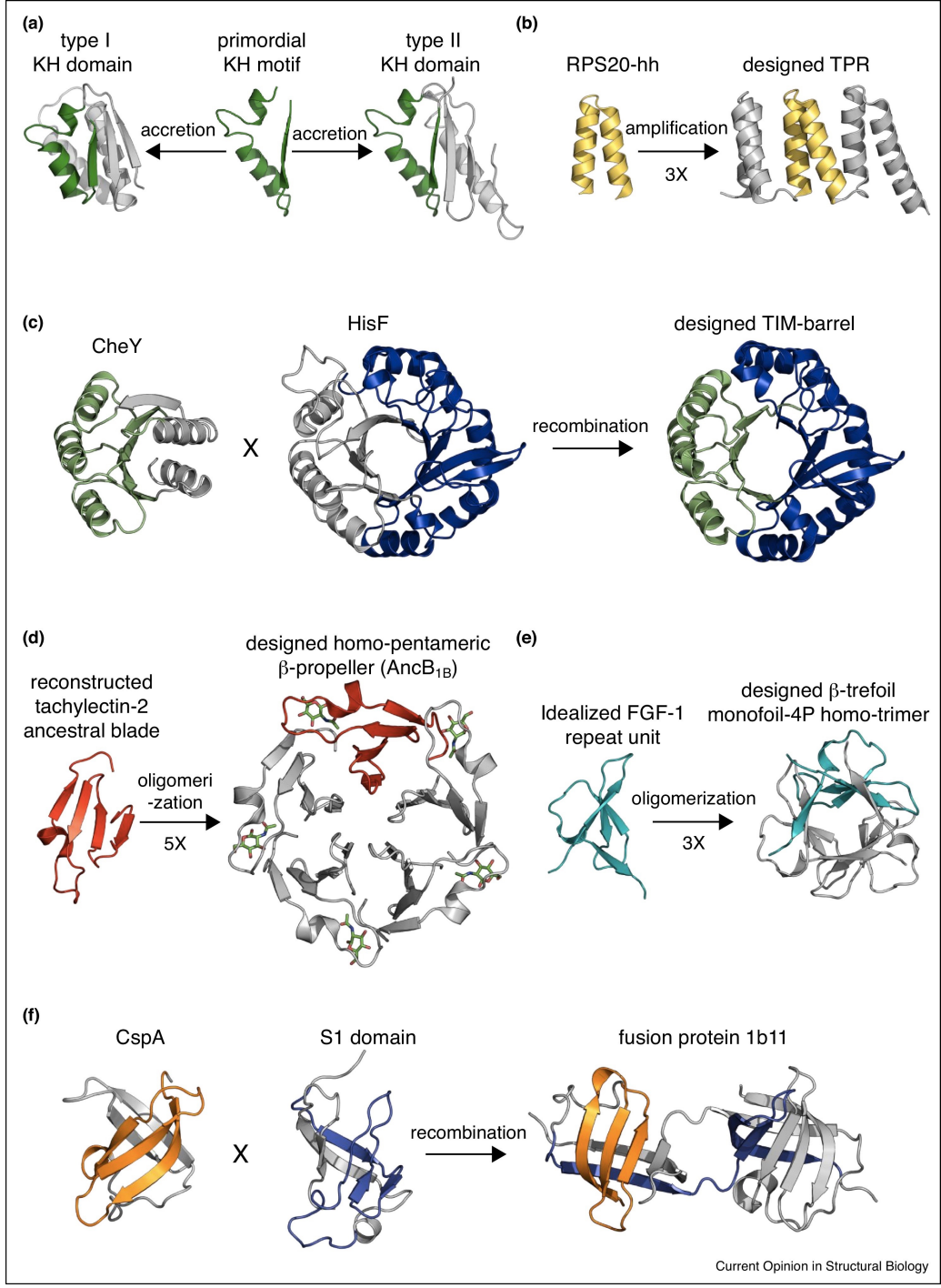


Transition from disorder to fold: TPR unit (Zhu et al., 2016)

- **Conclusions**:
  - folded proteins can emerge from conditionally folded proteins via neutral evolution
  - repetition is a fruitful way to create folded proteins

Same principles can be used in protein design

- Repetition is frequent
- Recombination is rare and difficult
- They are not mutually exclusive



Summary and possible exam questions

- HMMs for alignment of two sequences
- Profile HMMs for protein families
- Sequence similarity search using profile HMMs
- Relationship between structural similarity of proteins, their common evolutionary origin and distant sequence similarity detection using HMMs