Lesson 12.
Content

1. Neural networks.

2. Prediction of secondary structure.

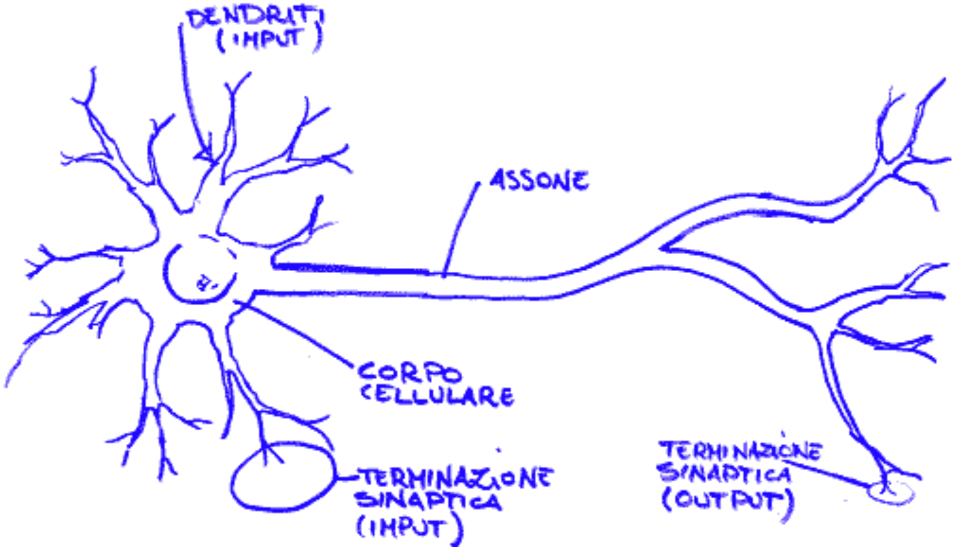
3. Protein contact prediction.

4. 3D structure prediction with Deep Learning.

Artificial neural networks (ANNs)

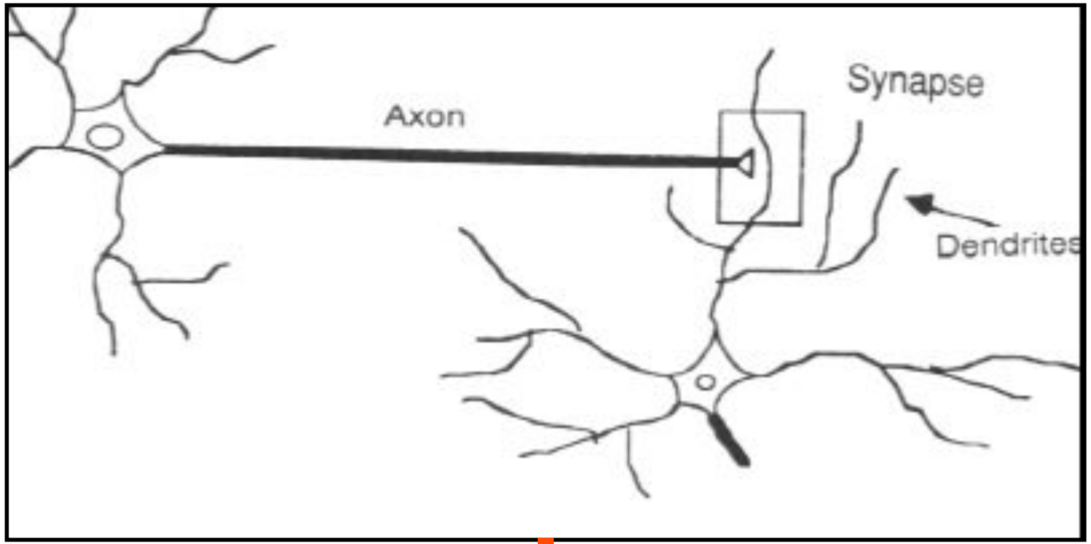
- Computational structures based on (inspired to) the anatomy and physiology of biological neural networks
- Initially developed to simulate information processing and learning in brain

Physiologically, a neuron receives excitatory and inhibitory stimuli (input) and emits a response signal (output) in case the intensity of the stimulus overcomes a given threshold

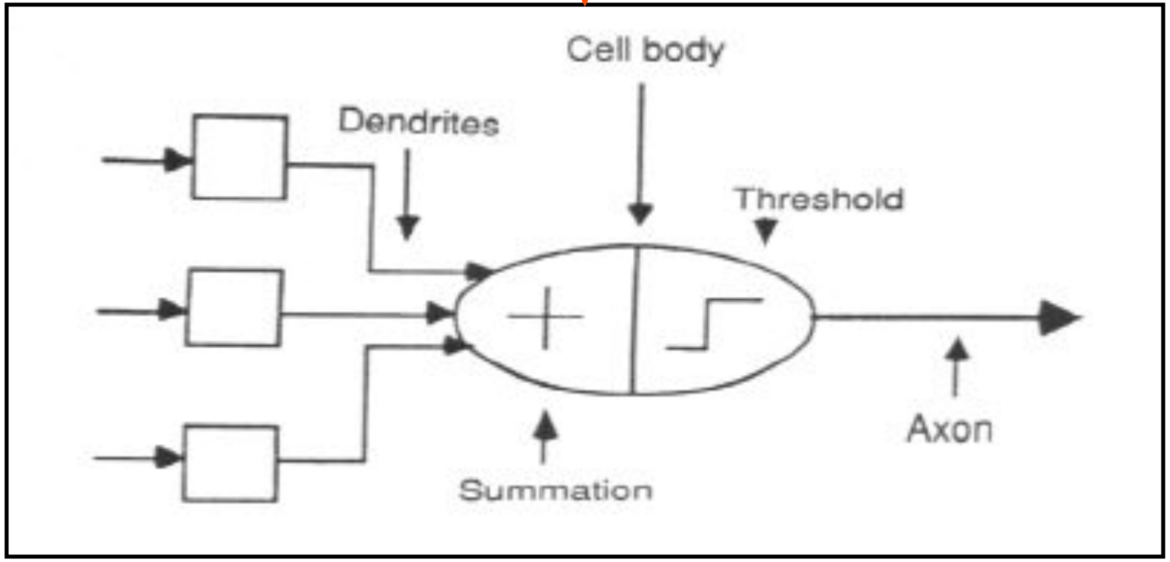


Artificial neural networks (ANNs)

synapse



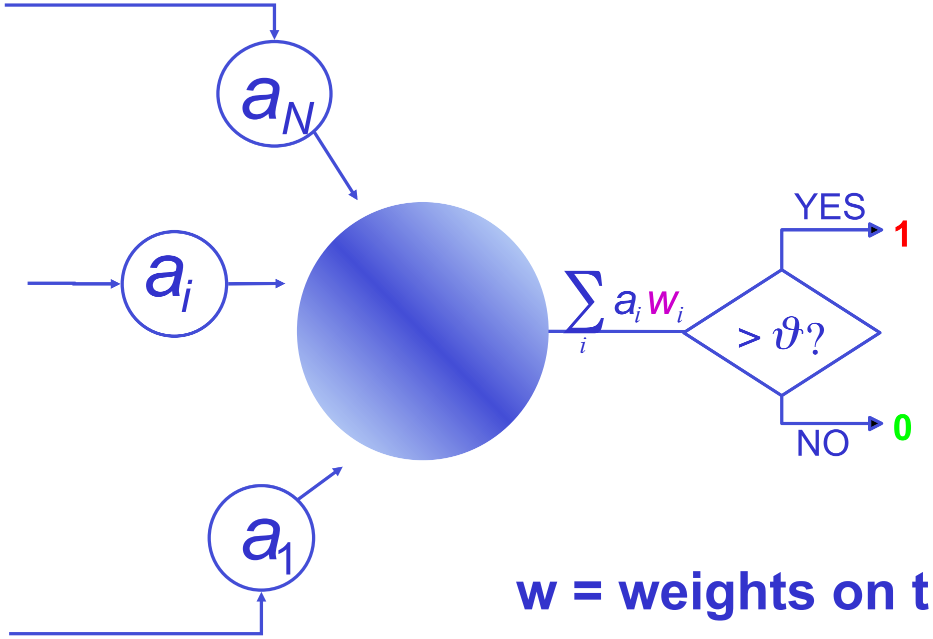
ANN



Artificial neural networks (ANNs)

- They are an example of machine-learning techniques, whose aim is automatically fitting a model value to a known value as closely as possible

- The algorithm will learn from a set of known examples by iterative changes to its parameters – weights of the input data – until the prediction best fits the reality



w = weights on the input data to be summed up

Artificial neural networks (ANNs)

- ANNs operate by processing information through "**layers**"; each layer can have many **nodes** or **units**

- The simplest NN is a two-layered network, an input layer and an output layer, called ***perceptron***

- The firing of a node in a NN is simulated by assigning the binary values of ***1*** or ***0*** to its output; ***1*** is assigned when the weighted sum of inputs exceeds a predetermined **threshold value**

	a_1	a_2	$\text{output}_{\text{expected}}$
Ex. 1	1	0.3	1
Ex. 2	1	1	1
Ex. 3	0	0.8	0
Ex. 4	0.5	0.4	0

$$\sum_i a_i w_i > \theta \Longrightarrow \mathit{YES}\ (1) \qquad \sum_i a_i w_i < \theta \Longrightarrow \mathit{NO}\ (0)$$

One of the solutions: $w_1 = 1$, $w_2 = 0.5$, $\theta = 0.9$

Ex. 1: $a_1 * w_1 + a_2 * w_2 = 1*1 + 0.3*0.5 = 1.15\ (>0.9) \rightarrow 1$

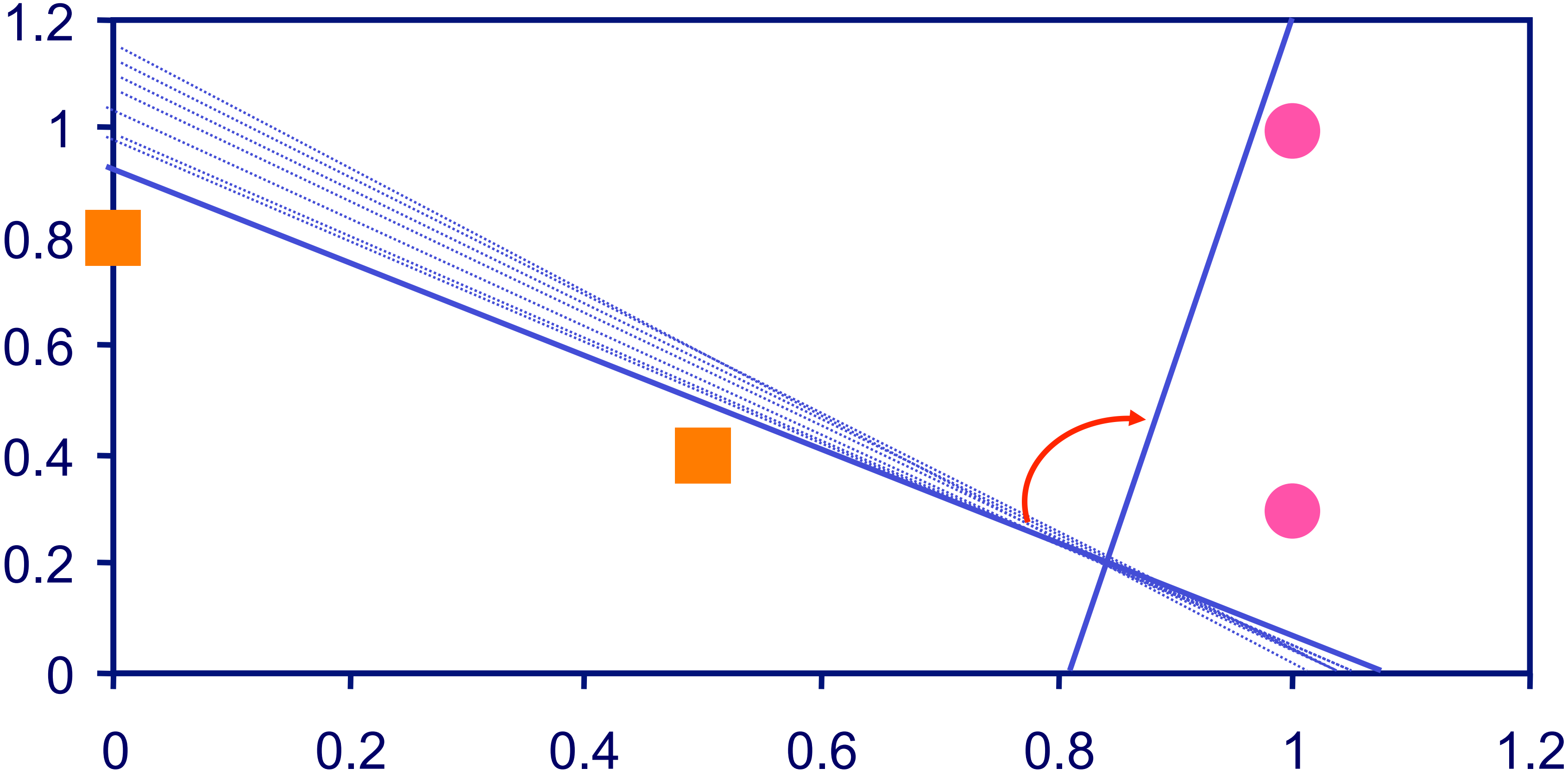
Ex. 2: $a_1 * w_1 + a_2 * w_2 = 1*1 + 1*0.5 = 1.5\ (>0.9) \rightarrow 1$

Ex. 3: $a_1 * w_1 + a_2 * w_2 = 0*1 + 0.8*0.5 = 0.4\ (<0.9) \rightarrow 0$

Ex. 4: $a_1 * w_1 + a_2 * w_2 = 0.5*1 + 0.4*0.5 = 0.7\ (<0.9) \rightarrow 0$

	x	y	$\text{output}_{\text{expected}}$
1	1	0.3	1
2	1	1	1
3	0	0.8	0
4	0.5	0.4	0

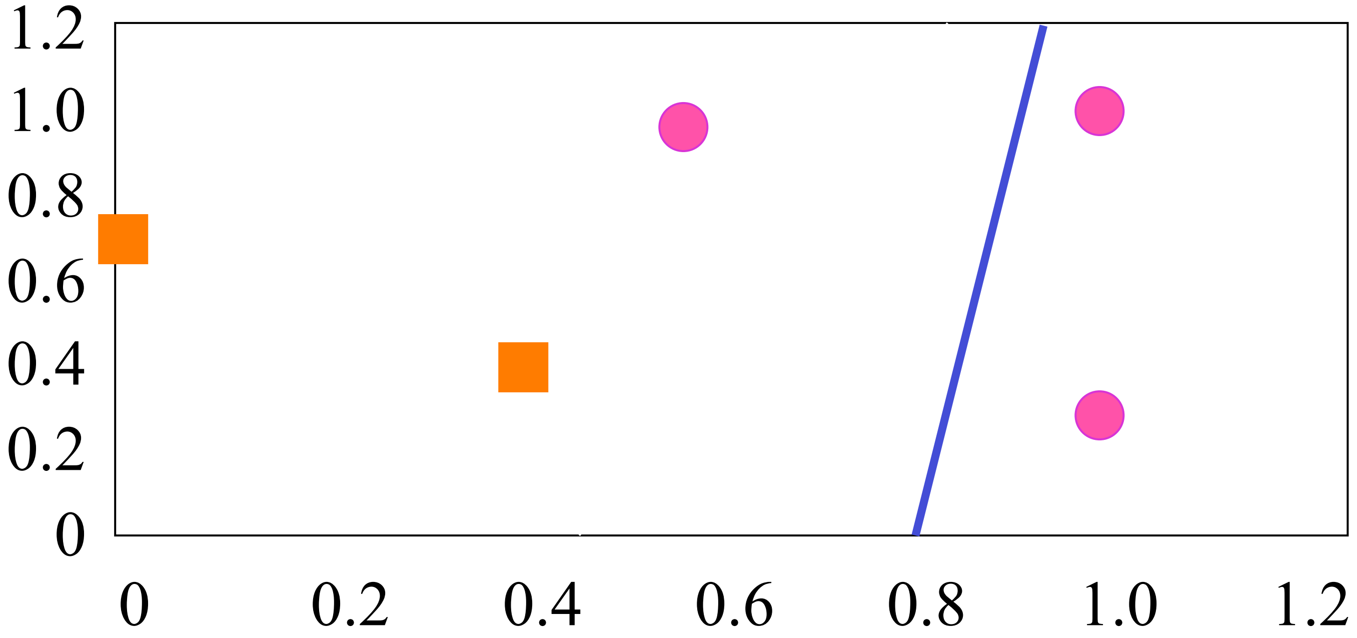
$$y = \underline{a}x + \underline{b}$$



Example of a 2D network

1) We assign 2 values (coordinates, a_i) to each point & associate a positive ● or negative ■ output

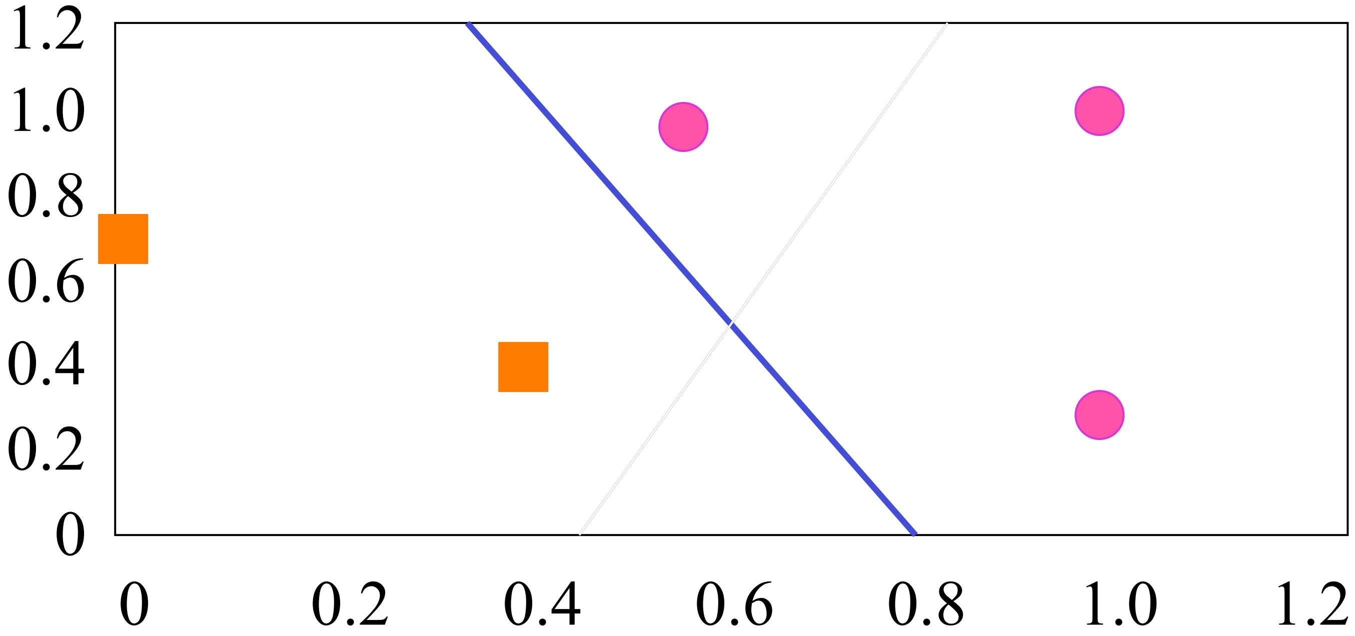
$$X=\Sigma a_i W_i$$



Example of a 2D network

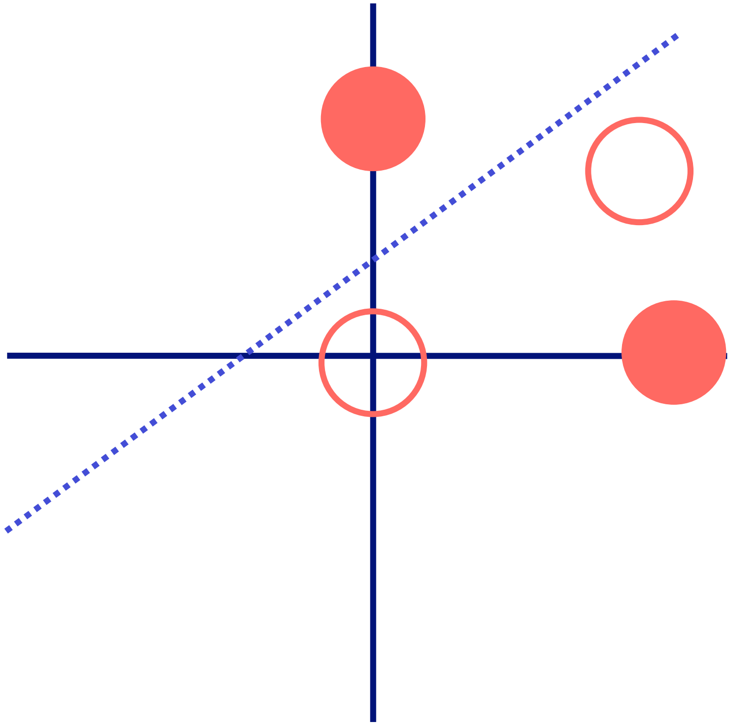
1) We assign 2 values (coordinates, a_i) to each point & associate a positive ● or negative ■ output

$$X=\Sigma a_i W_i$$



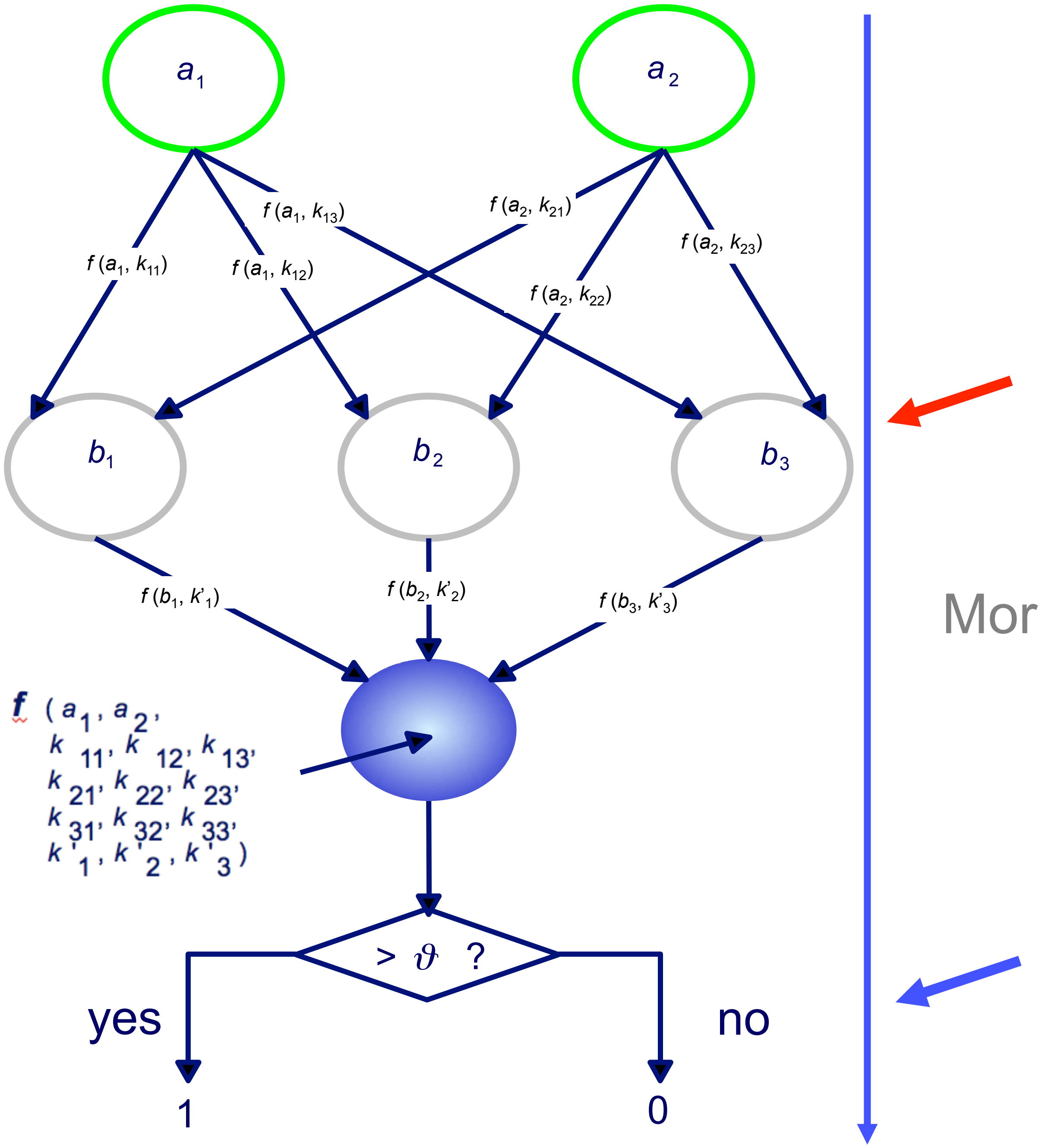
a neural networks can learn from its own mistakes

What function best discriminates between ● & ○ ?



A simple ANN would find at most a dashed straight line

➔ We need a more complex network, by introducing an *hidden layer*

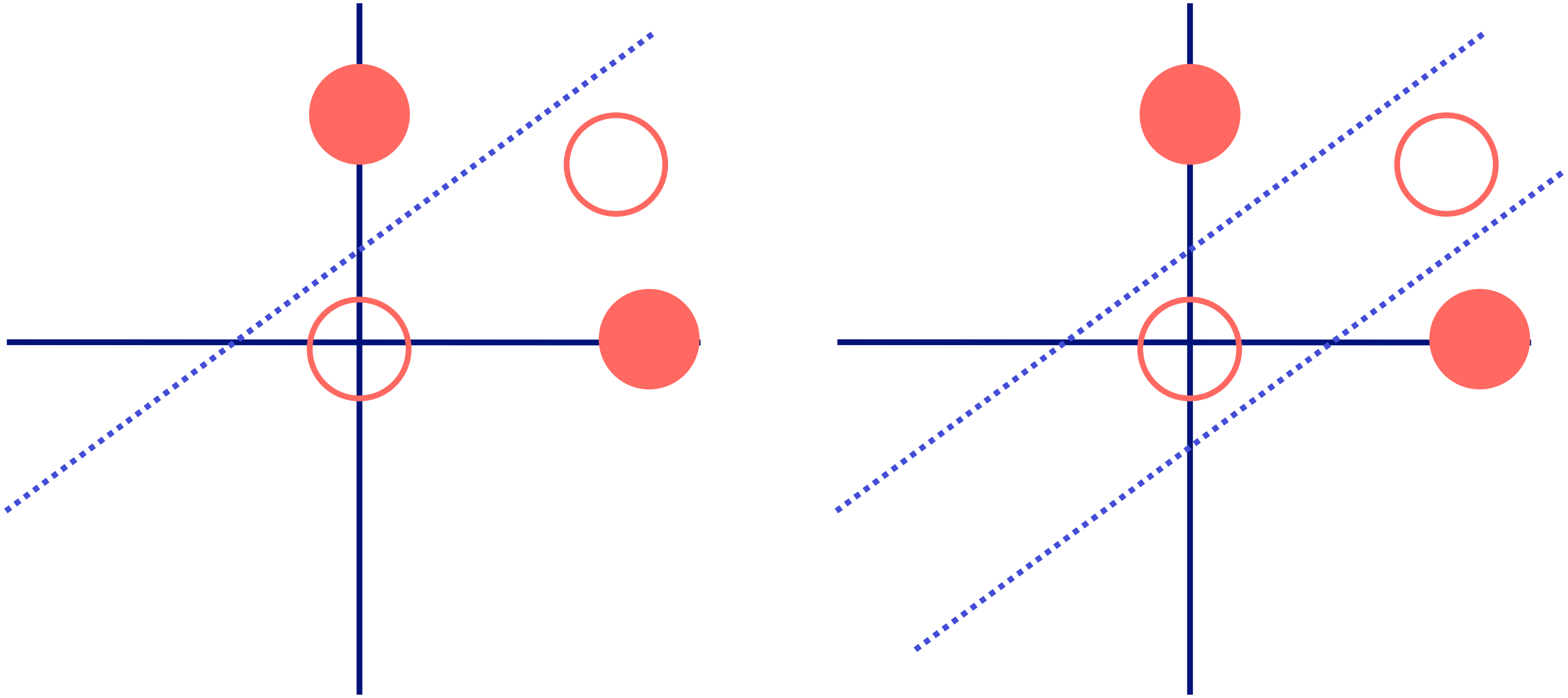


Hidden layer

More parameters to be optimized

Feed-forward ANN (direction)

Solution

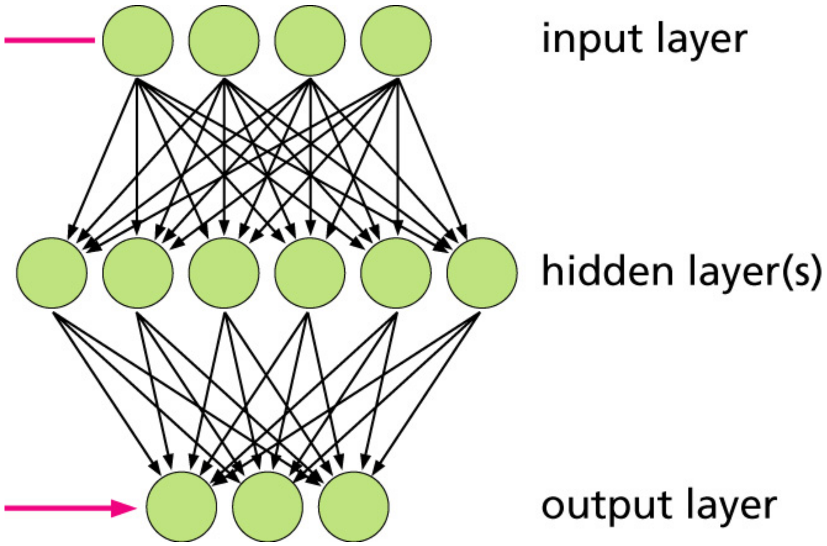


A straight line is not enough to solve the problem, we need two!

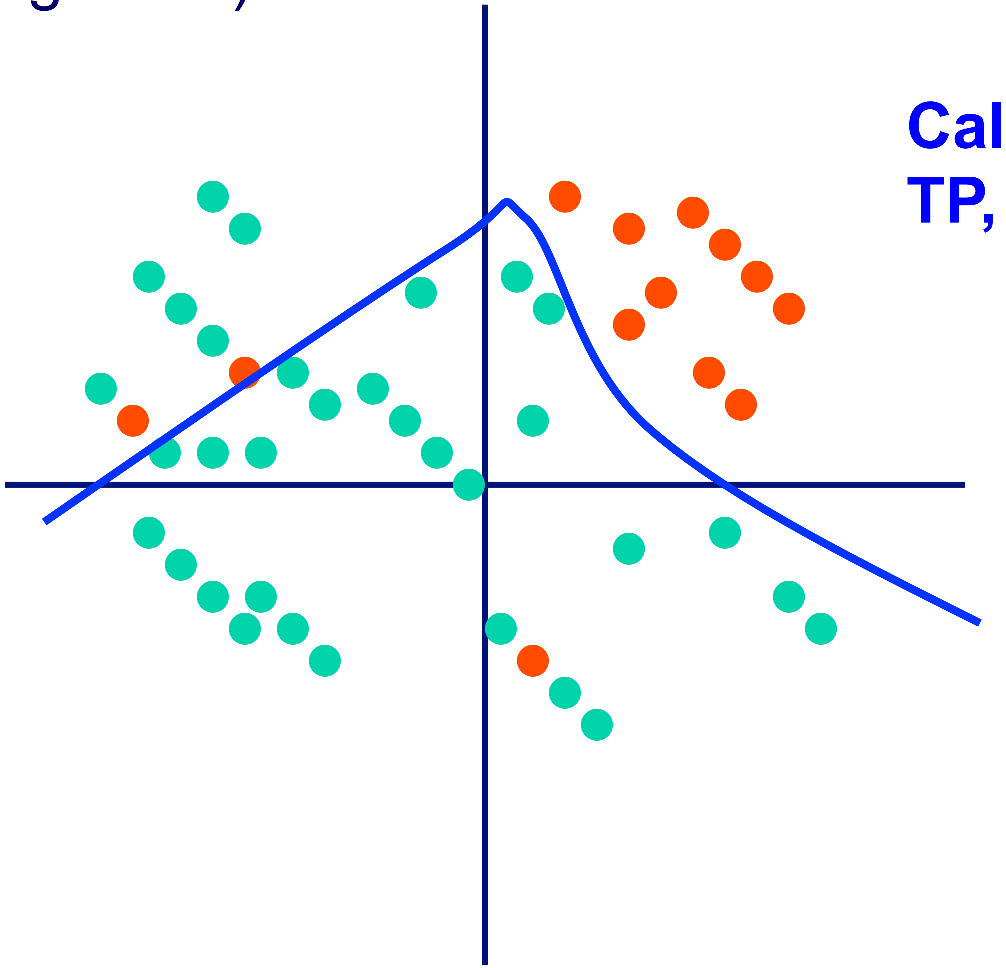
Artificial neural networks (ANNs)

- A more complex and more common NN is one that has one or more layers between the input and output ones, the so-called **hidden layers**

- The hidden layers perform nonlinear transformations of the inputs entered into the network, because there is more than one path to the output node



What function best discriminates between ● (positives) & ● (negatives)?



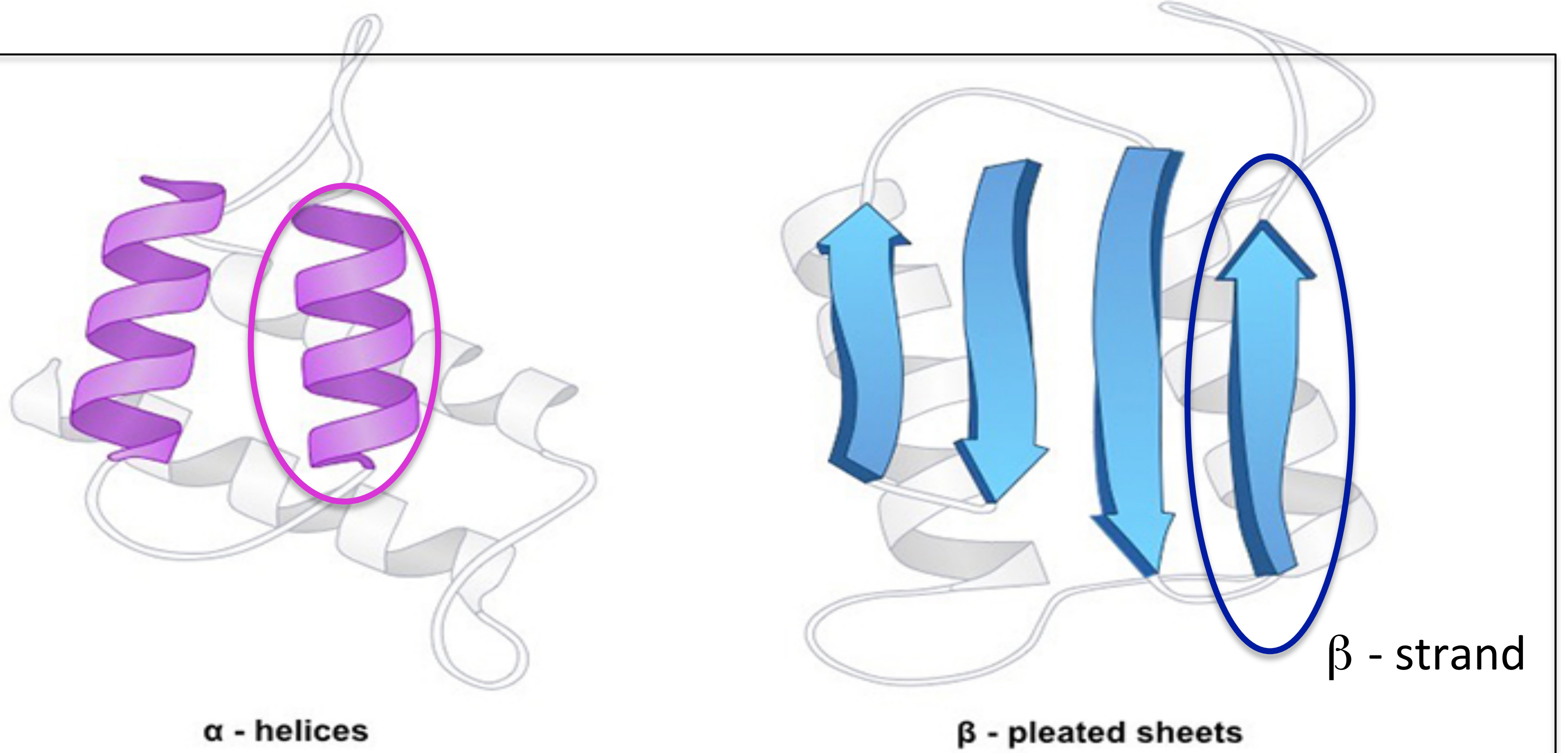
Calculation of: TP, FP, TN, FN

No matter how sophisticated the network is, it will always generate some incorrect predictions (FP & FN)

All statistical methods need a validation to be confidently used

ANNs for the prediction of secondary structure (SS)

- NNs have been widely used in Bioinformatics for the prediction of the secondary structure (SS) of proteins

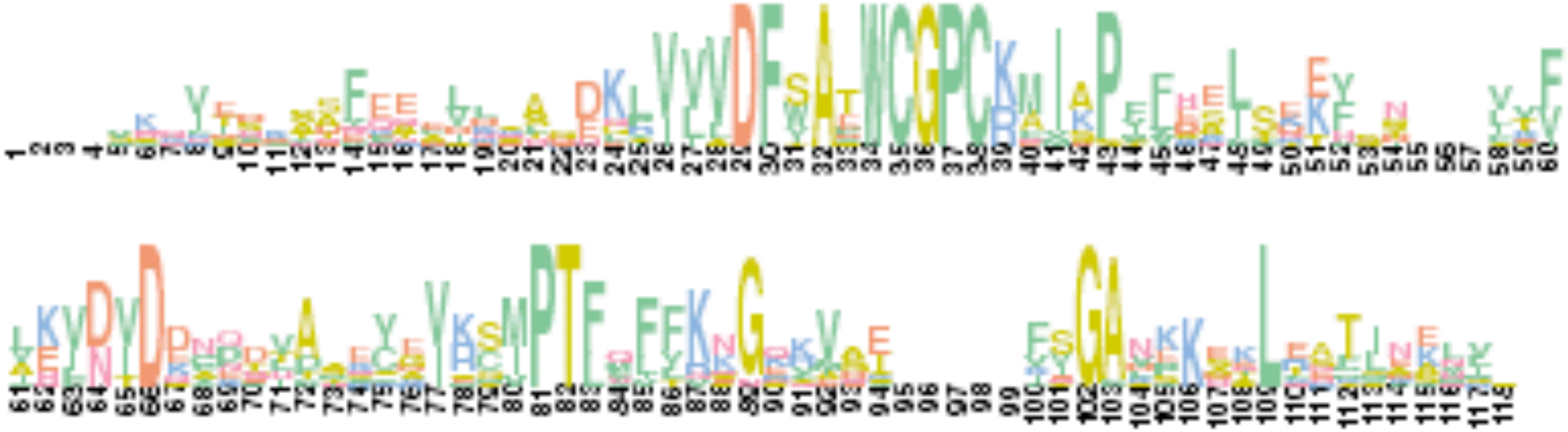


α-helices and β-strands are the only regular protein secondary structure motifs; they are connected by turns (ordered 3/4-residue motifs) or loops

ANNs for the prediction of secondary structure (SS)

- Application of NNs to the prediction of the protein secondary structure is ideal for at least two reasons:

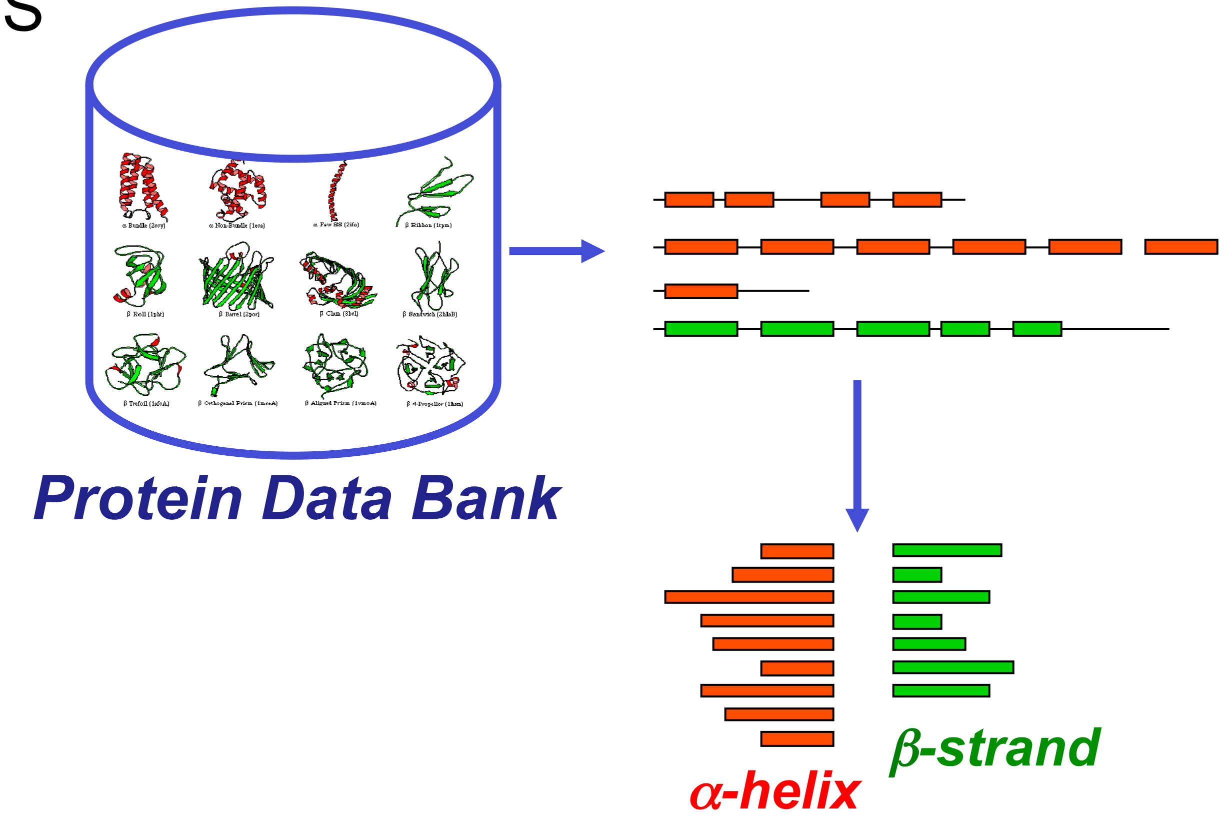
1. The NN prediction is context-dependent, i.e. different positions in the sequence (or alignment) can have a different relevance (weight) for the prediction



ANNs for the prediction of secondary structure (SS)

- Application of NNs to the prediction of the protein secondary structure is ideal for at least two reasons:

2. Many examples to learn from are available for the protein SS

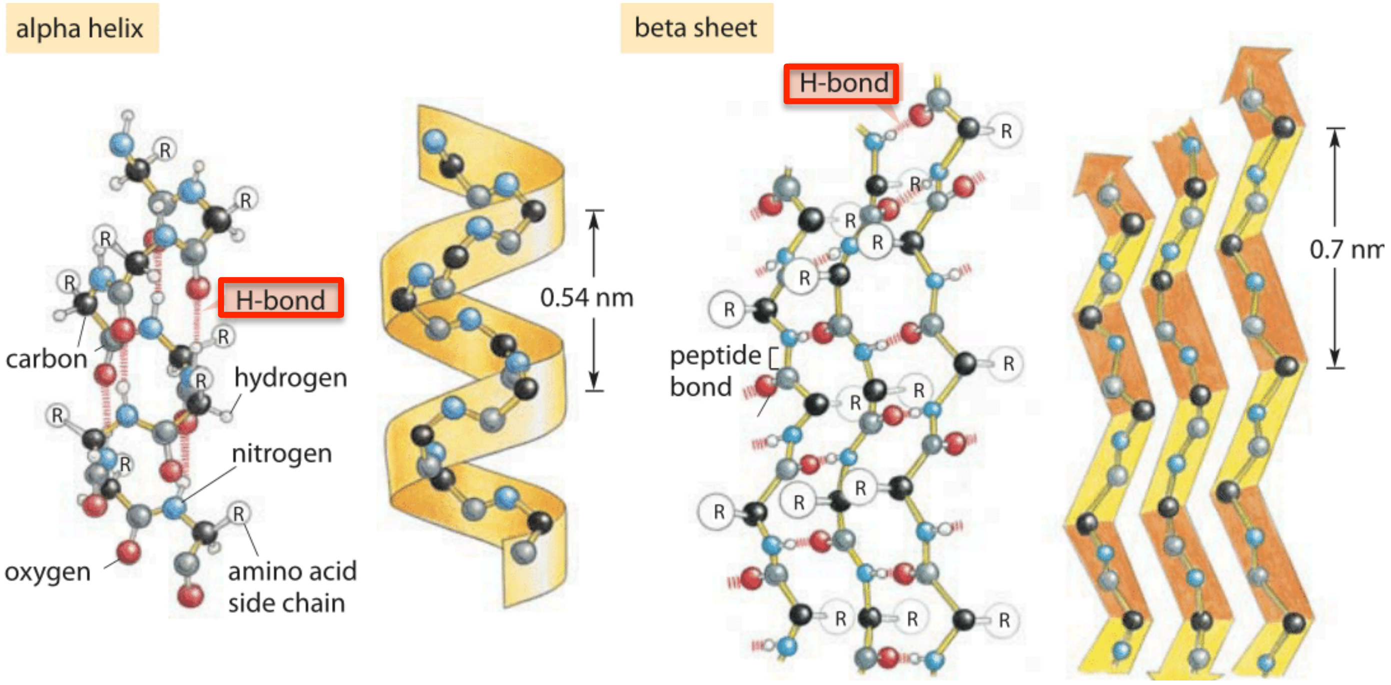


Defining the protein SS: DSSP (*Dictionary of protein secondary structure)*

http://swift.cmbi.umcn.nl/gv/dssp/

DSSP-software: assigns the SS according to hydrogen-bond patterns

DSSP-database: contains SS assignments (plus more info) for all the protein entries in the PDB.



Defining the protein SS: DSSP (*Dictionary of protein secondary structure)*

The DSSP code

- **H** = alpha helix
- **B** = residue in isolated beta-bridge
- **E** = extended **strand**, participates in **beta-sheet**
- G = 3-helix (3/10 helix)
- **I** = 5 helix (pi helix)
- **T** = hydrogen bonded turn
- **S** = bend
- Blank = loop or irregular

```
Sequence:    MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLT-SLDAAKSELDKAIGRNTNGV
DSSP:            HHHHHHHHH EEEEEE TTS EEEETTEE - HHHHHHHHHHHHTS TTB
Sequence:    ITKDEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVRRAALINMVFQMGETGVA
DSSP:          HHHHHHHHHHHHHHHHHHHHH TTTHHHHHHS HHHHHHHHHHHHHHHHHHHH
Sequence:    GFTNSLRMLQQKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDAYK
DSSP:        T HHHHHHHHTT HHHHHHHHHSSHHHHHSHHHHHHHHHHHHHSSSGGG
```

PDB ID: 103L (hydrolase)

ANNs for the prediction of secondary structure (SS)

- The input signal for an amino acid is usually a group of 20 units in the input layer; the signals of the input will be all ***0*** except that representing the particular residue, which will be ***1***
- Usually the sequence is sampled by a sliding window, with the central residue being that for which the SS is predicted (the input is thus a long string of ***0***/***1***: for a 13-res window 13x20 units)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	Y	W
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
I	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
C	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

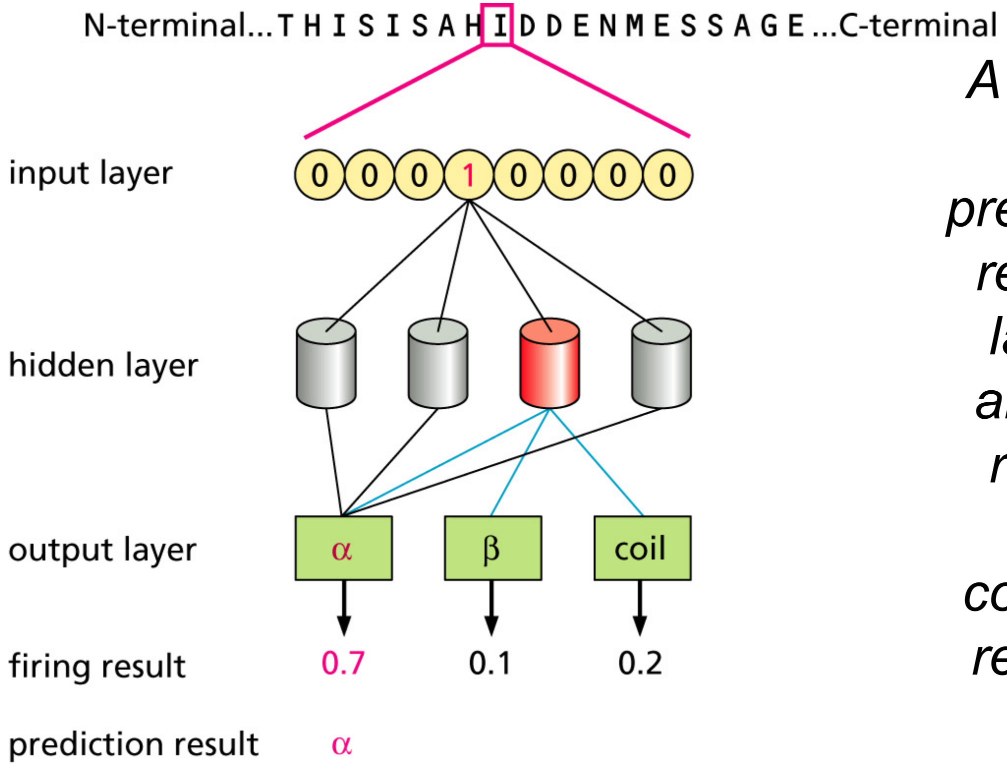
ANNs for the prediction of secondary structure (SS)

- When using multiple aligned sequences, the input layer signals will be related to **sequence profiles** based on these alignments
- Information contained in multiple alignments increases the accuracy of prediction, because proteins preserve their SS during evolution

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	Y	W
A A A A V	.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.2	0	0
D D D E E	0	0	.6	.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E E E E E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F F W F F	0	0	0	0	.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.2
I L L L L	0	0	0	0	0	0	0	.2	0	.8	0	0	0	0	0	0	0	0	0	0
G G G G G	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H H H H H	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
C C C C C	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I L L L L	0	0	0	0	0	0	0	.2	0	.8	0	0	0	0	0	0	0	0	0	0
K K R K R	0	0	0	0	0	0	0	0	.6	0	0	0	0	0	.4	0	0	0	0	0
L I I I I	0	0	0	0	0	0	0	.8	0	.2	0	0	0	0	0	0	0	0	0	0
M M M M M	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
N N S S S	0	0	0	0	0	0	0	0	0	0	0	.4	0	0	.6	0	0	0	0	0
C C C C C	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

ANNs for the prediction of secondary structure (SS)

- The output layer usually consists of 3 units, corresponding to the three alternative conformations to predict (a-helix, b-strand, loop/coil)
- An output like (1, 0, 0) would correspond to a perfect helix prediction; however prediction is usually done based on the highest number in output (see below)



A simplified representation of a multilayer NN: prediction is made on the central residue of the window (an Ile*); layer nodes receiving signals above a certain value, e.g. the red one, will fire to the output layer (prediction: helix); confidence of prediction can be related to how close to 1 is the highest number*

ANNs for the prediction of secondary structure (SS): procedure

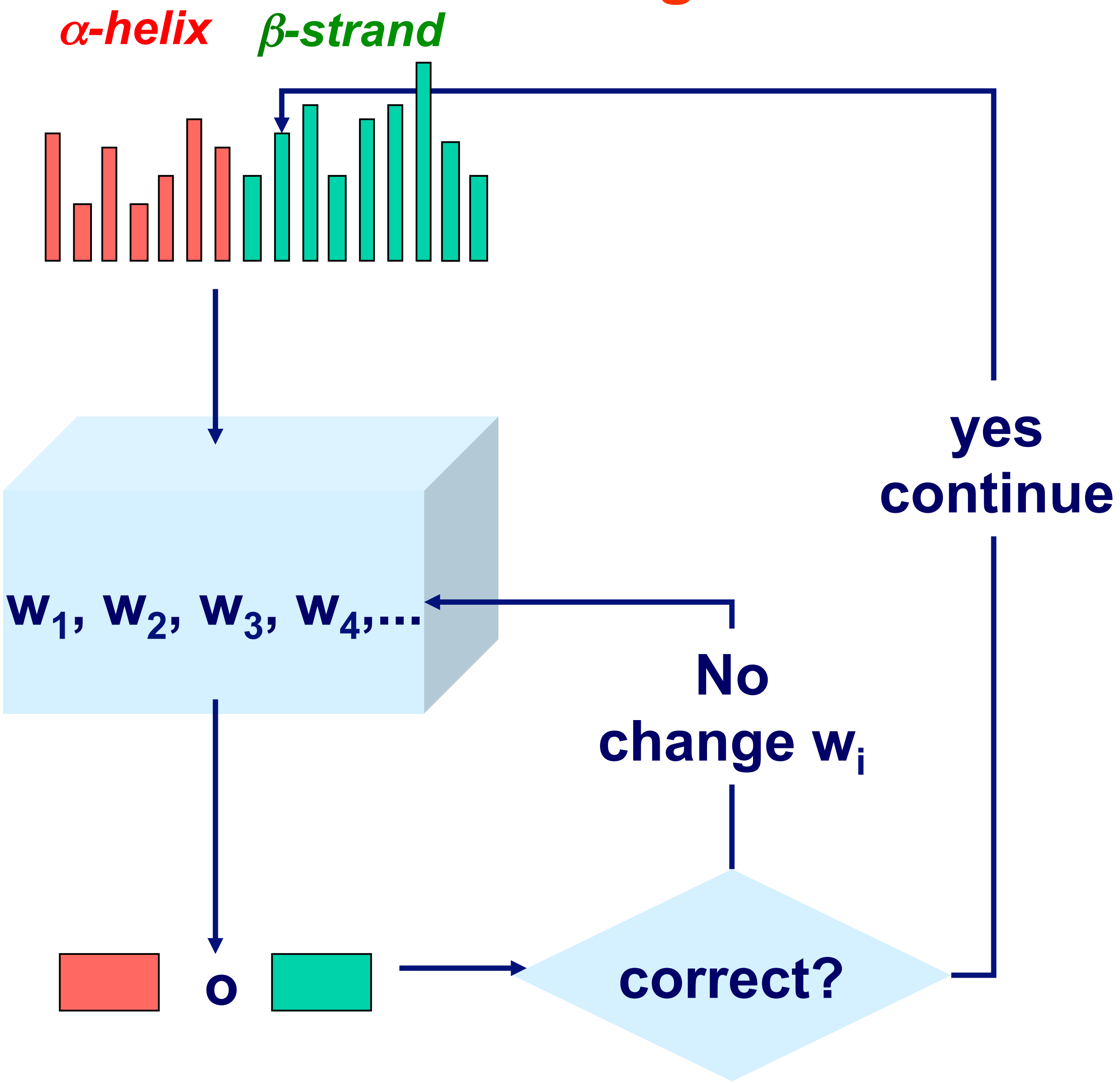
Implementing a NN requires three phases:

- *Training*: method development using non-homologous protein sequences of known structure

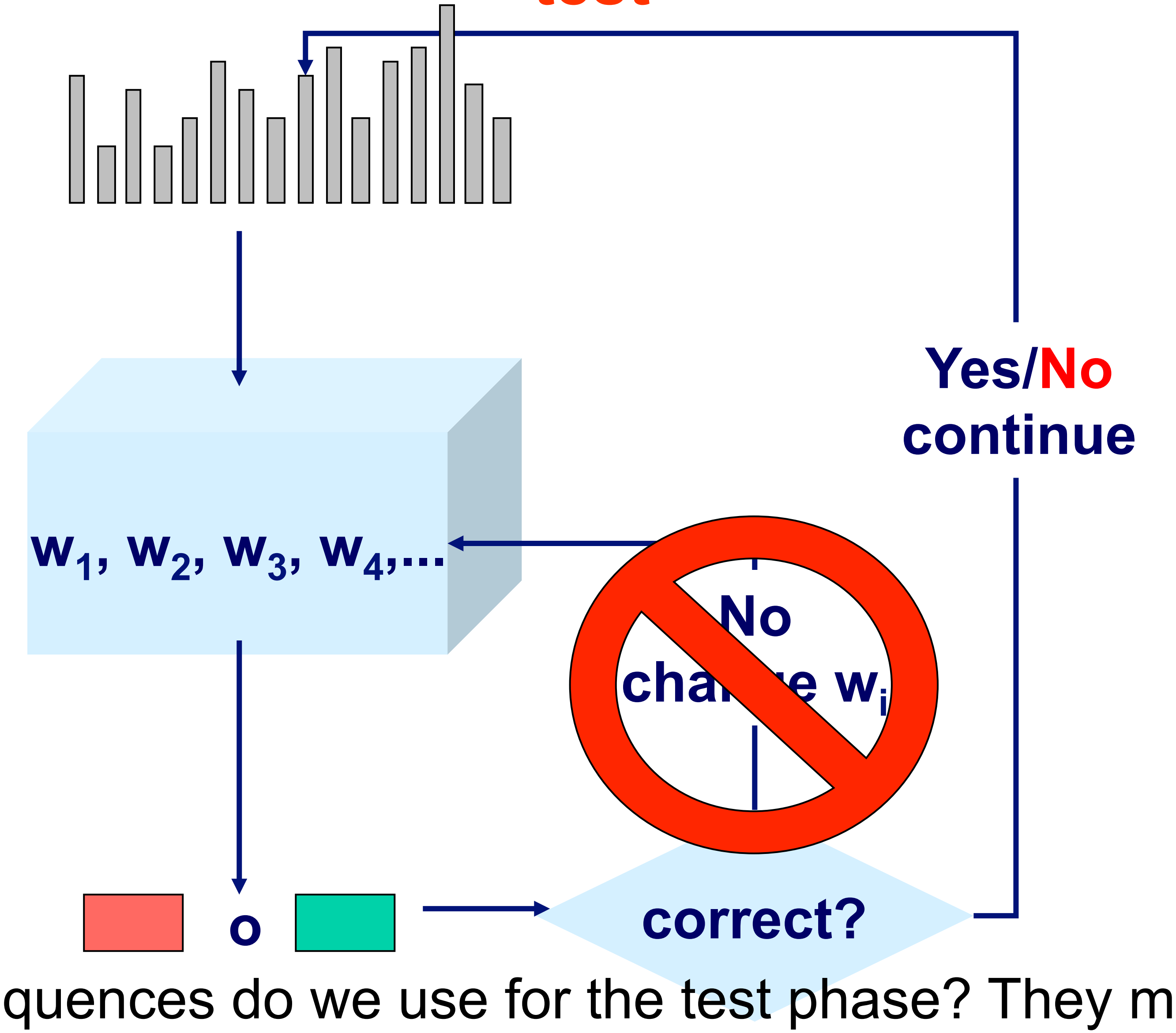
- *Test*: check of the method on protein sequneces of known structure

- *Validation*: statistical analysis of obtained results

ANNs for the prediction of secondary structure (SS): training



ANNs for the prediction of secondary structure (SS): test



What sequences do we use for the test phase? They must also be structure-known

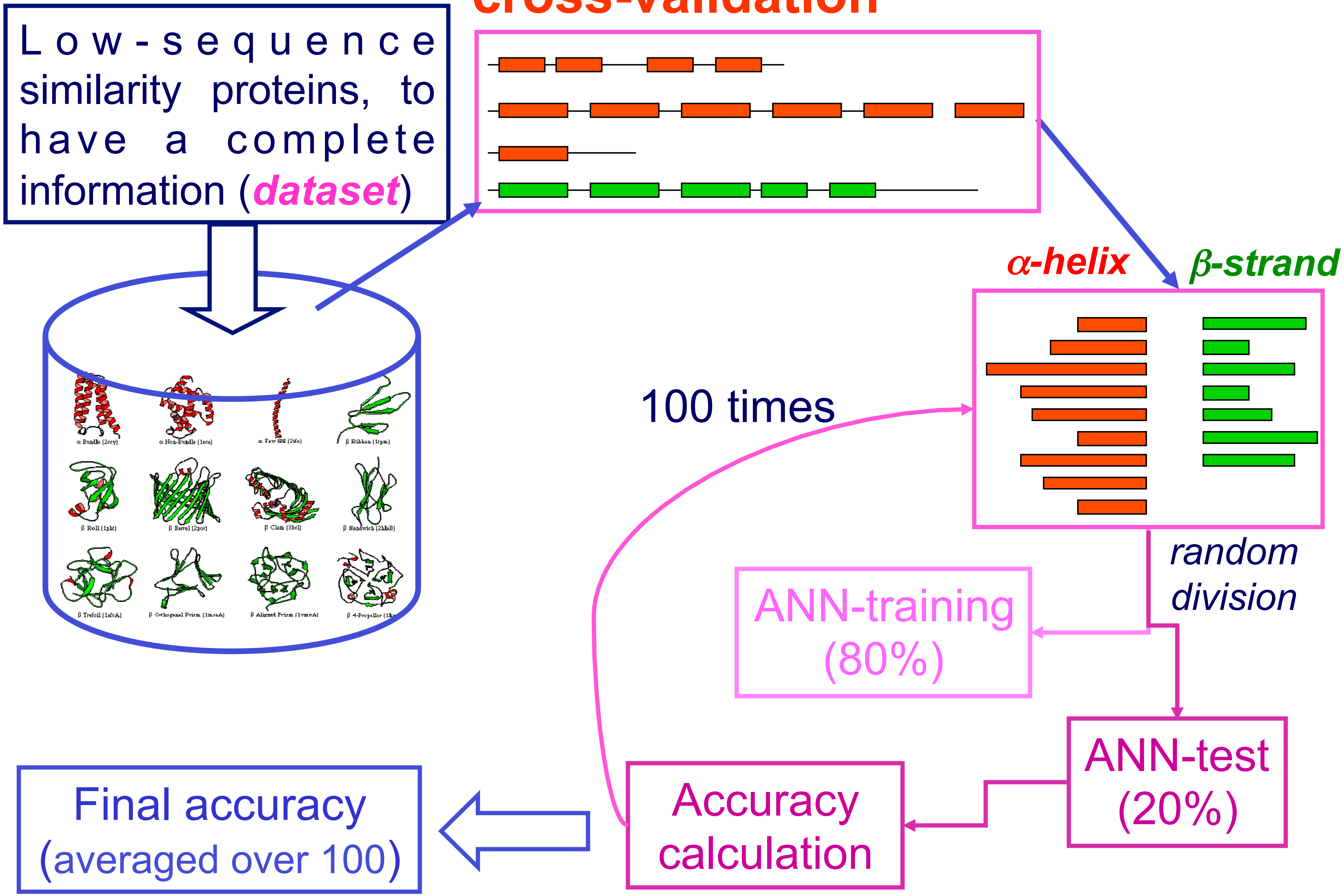
Validation

- To be reliable, knowledge-based methods must be tested with a rigorous statistics

- The most commonly used validation statistics is the **cross-validation** (or *jack-knife test*)

- From cross-validation results measures of the prediction performance (such as sensitivity, specificity, correlation coefficient etc.) can be calculated, which are universal, therefore comparable and reproducible

ANNs for the prediction of secondary structure (SS): cross-validation



Low-sequence similarity proteins, to have a complete information (*dataset*)

α-helix

β-strand

100 times

random division

ANN-training (80%)

ANN-test (20%)

Accuracy calculation

Final accuracy (averaged over 100)

Accuracy evaluation parameters

- **Q3** = percentage of sequence expected to have a correct SS prediction based on 3-state classification, H-E-C

$$Q3 = \left(\frac{TP(H)}{Tot(H)} + \frac{TP(E)}{Tot(E)} + \frac{TP(C)}{Tot(C)} \right) * 100$$

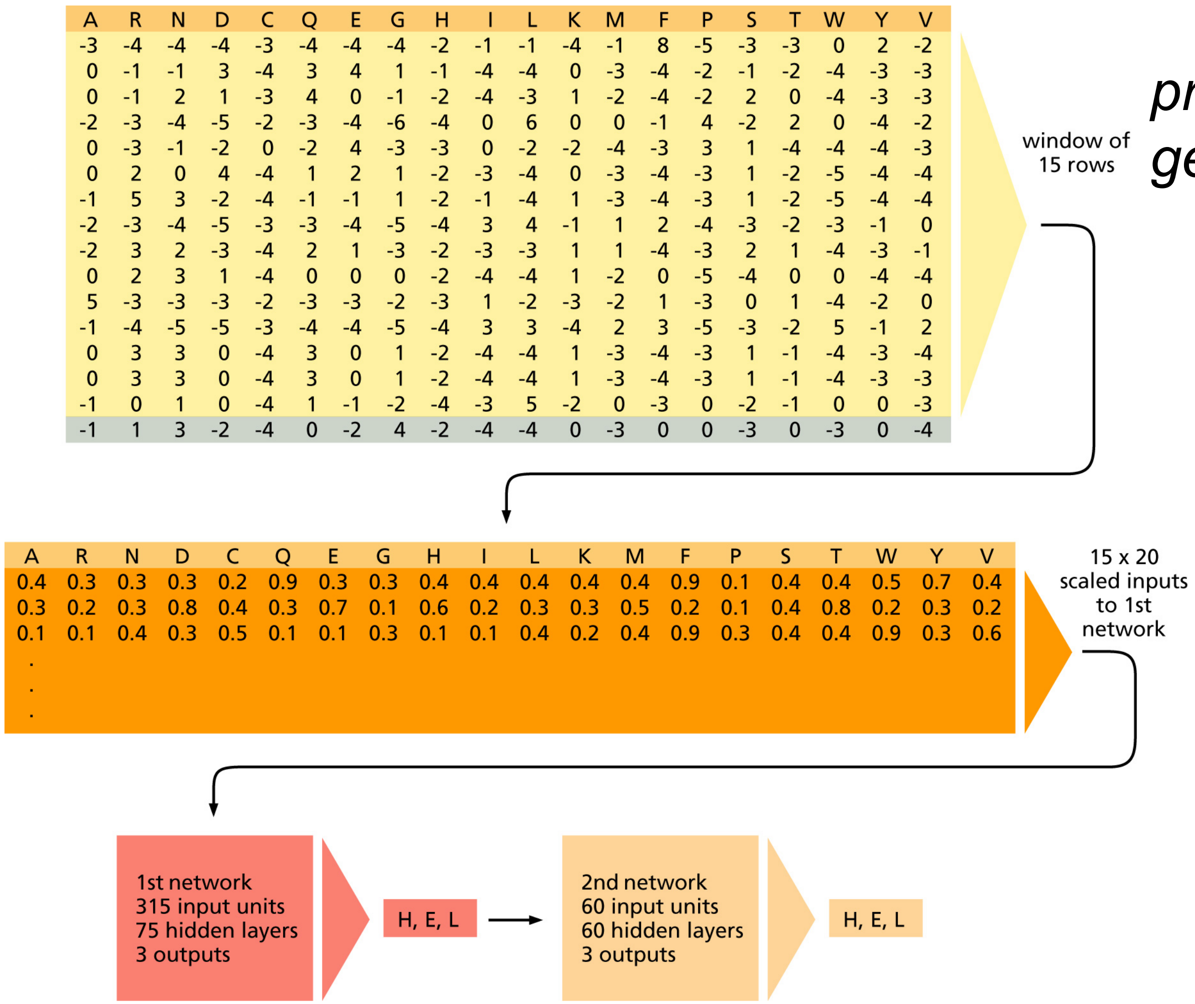
- **Matthew's** = geometrical mean of the correlation coefficients relative to the three states H-E-C (preferable to Q3)

$$CC_H = \frac{TP(H)*TN(H) - FP(H)*FN(H)}{\sqrt{(TP(H)+FP(H))*(TP(H)+FN(H))*(TN(H)+FP(H))*(TN(H)+FN(H))}}$$

$$CC_M = \sqrt[3]{CC_H * CC_E * CC_C}$$ *(geometrical mean)*

Server online: PSIPRED

Several SS prediction servers based on NNs are available, including **PSIPRED** and **PHDsec**



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

window of 15 rows

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
0.4	0.3	0.3	0.3	0.2	0.9	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.9	0.1	0.4	0.4	0.5	0.7	0.4
0.3	0.2	0.3	0.8	0.4	0.3	0.7	0.1	0.6	0.2	0.3	0.3	0.5	0.2	0.1	0.4	0.8	0.2	0.3	0.2
0.1	0.1	0.4	0.3	0.5	0.1	0.1	0.3	0.1	0.1	0.4	0.2	0.4	0.9	0.3	0.4	0.4	0.9	0.3	0.6
.																			
.																			
.																			

15 x 20 scaled inputs to 1st network

1st network
315 input units
75 hidden layers
3 outputs

H, E, L

2nd network
60 input units
60 hidden layers
3 outputs

H, E, L

PSIPRED dual network prediction: first a raw profile generated by PSI-BLAST is taken and scaled to a 0-1 range. A window of 15 elements is fed to the 1st network, which performs the initial SS prediction using various residue parameters. This initial prediction is fed into a 2nd NN where it is filtered to produce the final three-state SS prediction

An example…

>gi|15595724|ref|AAG03916.1| transcriptional regulator Dnr [Pseudomonas aeruginosa PA01]

MEFQRVHQQLLQSHHLFEPLSPVQLQELLASSDLV

NLDKGAYVFRQGEPAHAFYYLISGCVKIYRLTPEG

QEKILEVTNERNTFAEAMMFMDTPNYVATAQAVVP

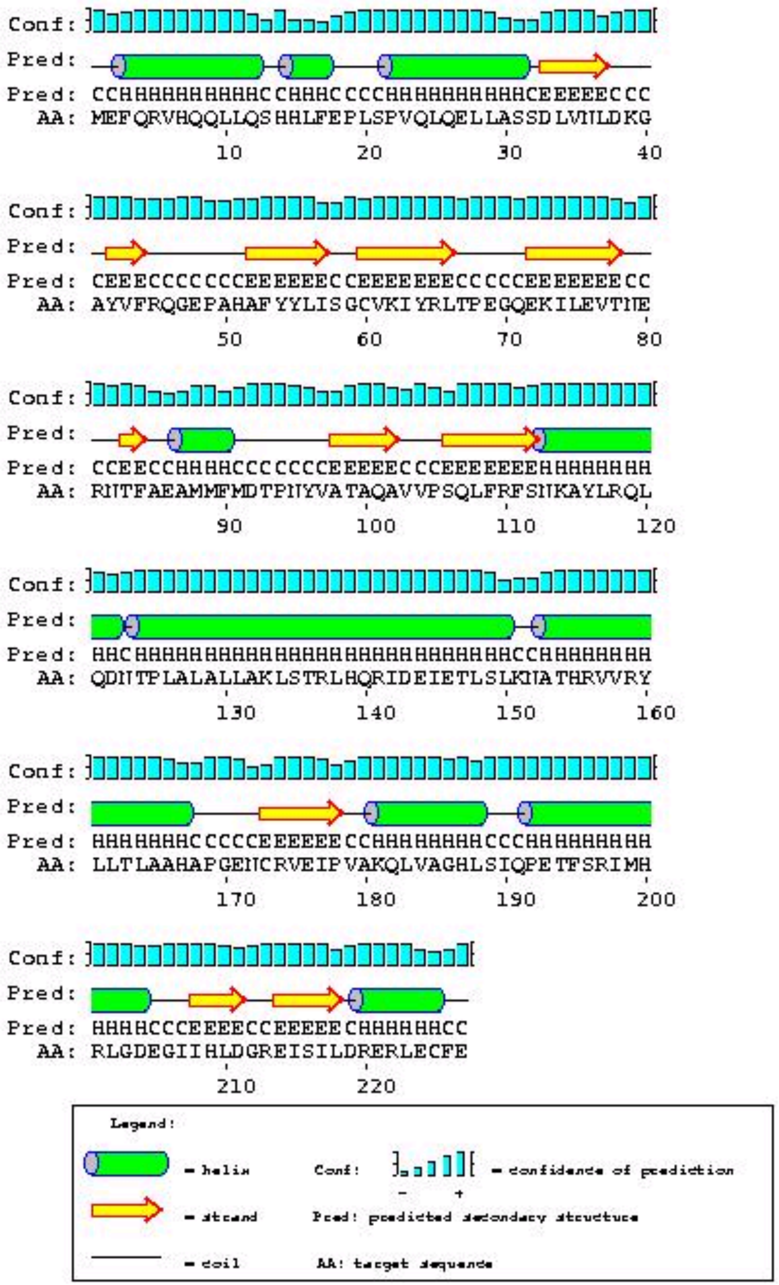
SQLFRFSNKAYLRQLQDNTPLALALLAKLSTRLHQ

RIDEIETLSLKNATHRVVRYLLTLAAHAPGENCRV

EIPVAKQLVAGHLSIQPETFSRIMHRLGDEGIIHL

DGREISILDRERLECFE

PSIPRED results



```
Conf:
Pred:
Pred: CCHHHHHHHHHHCCHHHCCCCHHHHHHHHHHCEEEEECCC
  AA: MEFQRVHQQLLQSHHLFEPLSPVQLQELLASSDLVNLDKG
              10        20        30        40

Conf:
Pred:
Pred: CEEECCCCCCCEEEEEEECCEEEEEEECCCCCEEEEEEEECC
  AA: AYVFRQGEPAHAFYYLISGCVKIYRLTPEGQEKILEVTNE
              50        60        70        80

Conf:
Pred:
Pred: CCEECCHHHHCCCCCCCEEEEECCCEEEEEEEEHHHHHHH
  AA: RNTFAEAMMFMDTPNYVATAQAVVPSQLFRFSNKAYLRQL
              90       100       110       120

Conf:
Pred:
Pred: HHCHHHHHHHHHHHHHHHHHHHHHHHHHCCHHHHHHHH
  AA: QDNTPLALALLAKLSTRLHQRIDEIETLSLKNATHRVVRY
             130       140       150       160

Conf:
Pred:
Pred: HHHHHHHCCCCCEEEEEECCHHHHHHHHCCCHHHHHHHHH
  AA: LLTLAAHAPGENCRVEIPVAKQLVAGHLSIQPETFSRIMH
             170       180       190       200

Conf:
Pred:
Pred: HHHHCCCEEEECCEEEEECHHHHHHCC
  AA: RLGDEGIIHLDGREISILDRERLECFE
             210       220
```

Legend:
- helix
- strand
- coil
Conf:
- confidence of prediction
-
+
Pred: predicted secondary structure
AA: target sequence

An example…

>gi|15595724|ref|AAG03916.1| transcriptional regulator Dnr [Pseudomonas aeruginosa PA01]

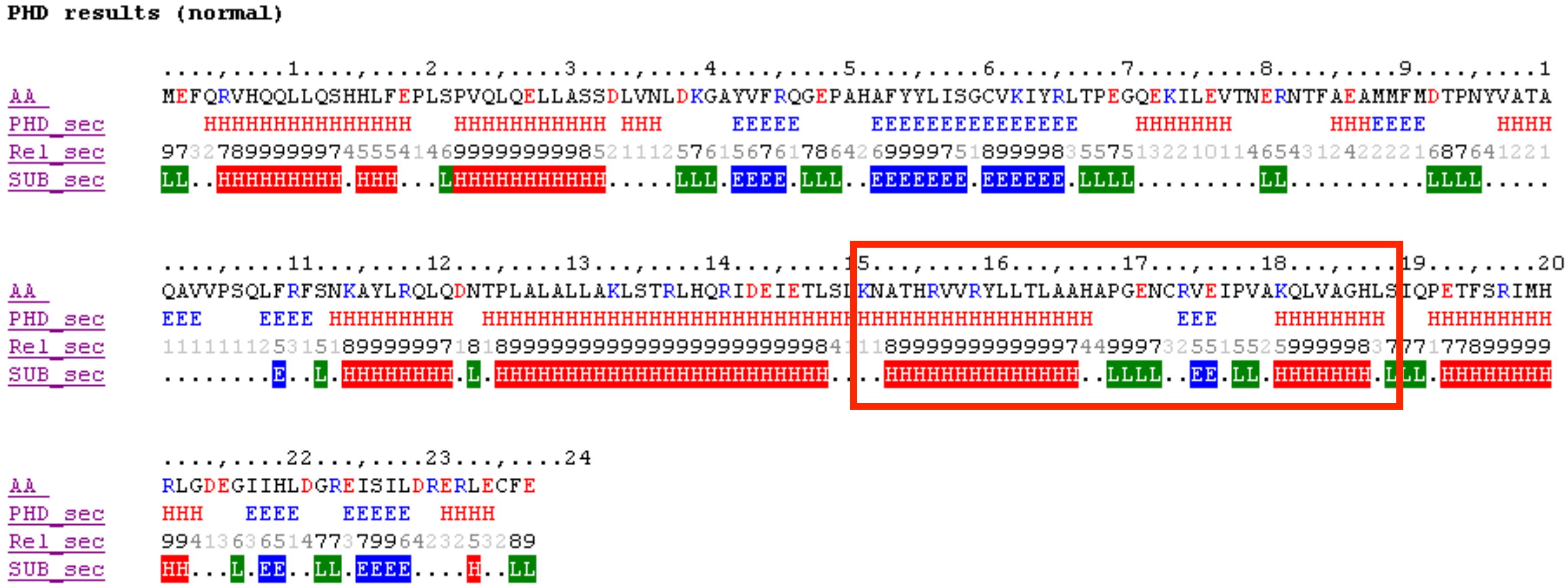
MEFQRVHQQLLQSHHLFEPLSPVQLQELLASSDLVNLDKGAYVFRQGEPAHAFYYLISGCVKIYRLTPEG

QEKILEVTNERNTFAEAMMFMDTPNYVATAQAVVPSQLFRFSNKAYLRQLQDNTPLALALLAKLSTRLHQ

RIDEIETLSLKNATHRVVRYLLTLAAHAPGENCRVEIPVAKQLVAGHLSIQPETFSRIMHRLGDEGIIHL

DGREISILDRERLECFE

PHDsec results

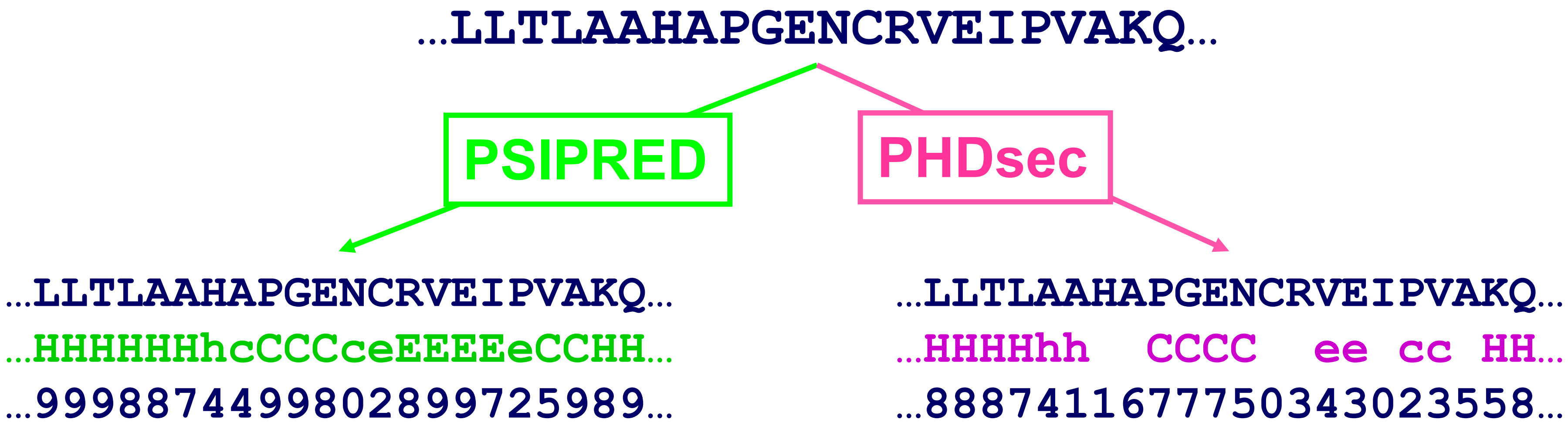


```
PHD results (normal)

            ....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8....,....9....,....1
AA          MEFQRVHQQLLQSHHLFEPLSPVQLQELLASSDLVNLDKGAYVFRQGEPAHAFYYLISGCVKIYRLTPEGQEKILEVTNERNTFAEAMMFMDTPNYVATA
PHD_sec        HHHHHHHHHHHHHHH   HHHHHHHHHHH HHH     EEEEE     EEEEEEEEEEEEEEE    HHHHHHH       HHHEEEE     HHHH
Rel_sec     9732789999997455541469999999999852111257615676178642699997518999983557513221011465431242222168764 1221
SUB_sec     LL..HHHHHHHHH.HHH...LHHHHHHHHHHH.....LLL.EEEE.LLL..EEEEEEE.EEEEEE.LLLL.........LL..........LLLL.....

            ....,....11...,....12...,....13...,....14...,....15...,....16...,....17...,....18...,....19...,....20
AA          QAVVPSQLFRFSNKAYLRQLQDNTPLALALLAKLSTRLHQRIDEIETLSLKNATHRVVRYLLTLAAHAPGENCRVEIPVAKQLVAGHLSIQPETFSRIMH
PHD_sec     EEE    EEEE HHHHHHHHH  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH      EEE    HHHHHHHH   HHHHHHHHH
Rel_sec     1111111253151899999971818999999999999999999999998411189999999999999744999732551552599999837771778999 99
SUB_sec     ........E..L.HHHHHHHH.L.HHHHHHHHHHHHHHHHHHHHHHHH....HHHHHHHHHHHHHH..LLLL..EE.LL.HHHHHHH.LLL.HHHHHHHH

            ....,....22...,....23...,....24
AA          RLGDEGIIHLDGREISILDRERLECFE
PHD_sec     HHH   EEEE   EEEEE  HHHH
Rel_sec     994136365147737996423253289
SUB_sec     HH...L.EE..LL.EEEE....H..LL
```

Confidence scores

To each predicted sequence position a confidence score is associated which indicates the probability of the prediction to be correct

Dnr da *Pseudomonas aeuruginosa*



...LLTLAAHAPGENCRVEIPVAKQ...

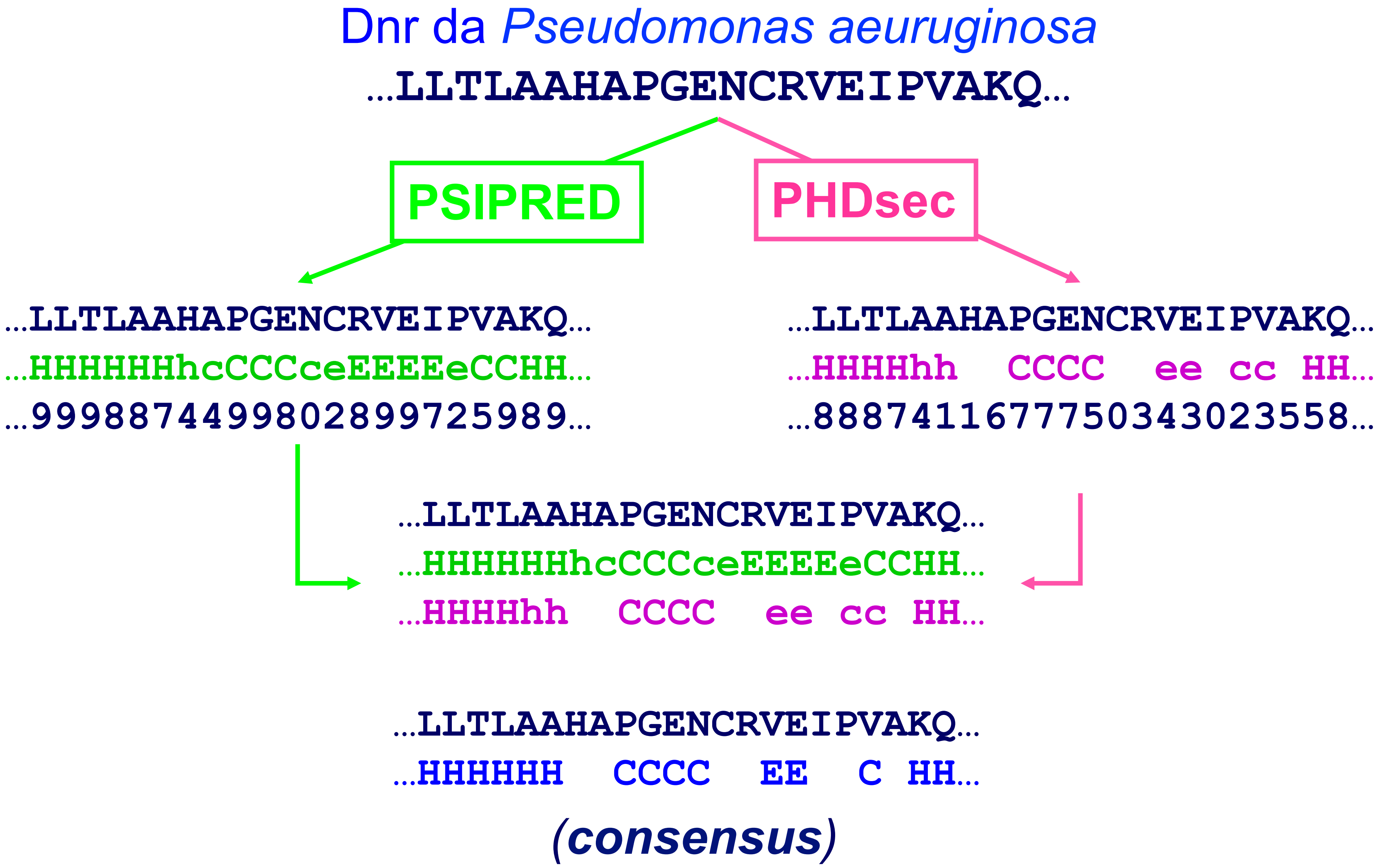
PSIPRED

PHDsec

```
...LLTLAAHAPGENCRVEIPVAKQ...
...HHHHHHhcCCCceEEEEeCCHH...
...9998874499802899725989...
```

```
...LLTLAAHAPGENCRVEIPVAKQ...
...HHHHhh  CCCC  ee cc HH...
...8887411677750343023558...
```

Metaserver: resources exploiting and combining the best SS prediction methods and improve their performance



Dnr da *Pseudomonas aeuruginosa*

```
...LLTLAAHAPGENCRVEIPVAKQ...
```

PSIPRED

PHDsec

```
...LLTLAAHAPGENCRVEIPVAKQ...
...HHHHHHhcCCCceEEEEeCCHH...
...9998874499802899725989...
```

```
...LLTLAAHAPGENCRVEIPVAKQ...
...HHHHhh  CCCC  ee cc HH...
...8887411677750343023558...
```

```
...LLTLAAHAPGENCRVEIPVAKQ...
...HHHHHHhcCCCceEEEEeCCHH...
...HHHHhh  CCCC  ee cc HH...
```

```
...LLTLAAHAPGENCRVEIPVAKQ...
...HHHHHH  CCCC  EE  C HH...
```

(consensus)

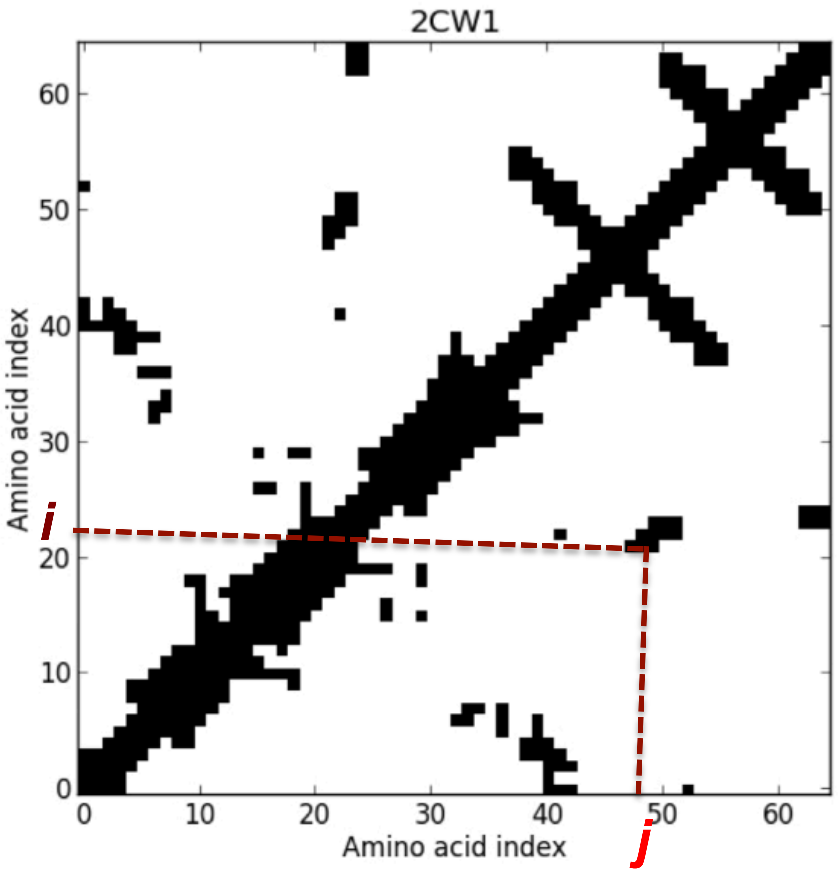
ANNs for the prediction of secondary structure (SS)

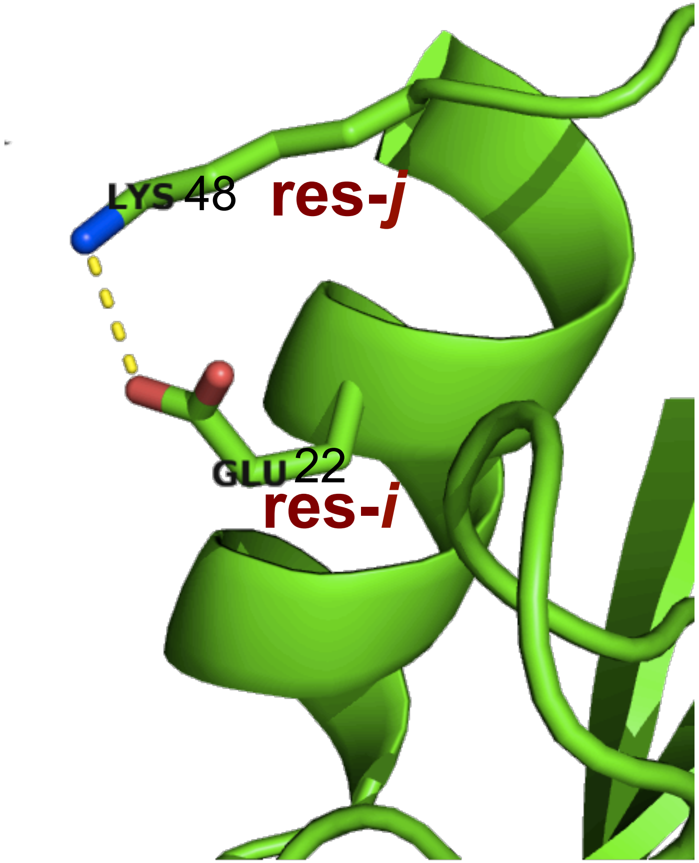
Accuracy of NN-based methods for the prediction of protein secondary structure can be vary high, up to 90-95%

Accuracy for a given query depends on the availability of homologs for it, i.e. on the availability of evolutionary information…

Protein contact map

A protein contact map is a 2D representation of a protein where a black dot is present at the cross-over of two residues (*i* and *j*), if they are closer than a given cut-off distance (usually 6 Å).

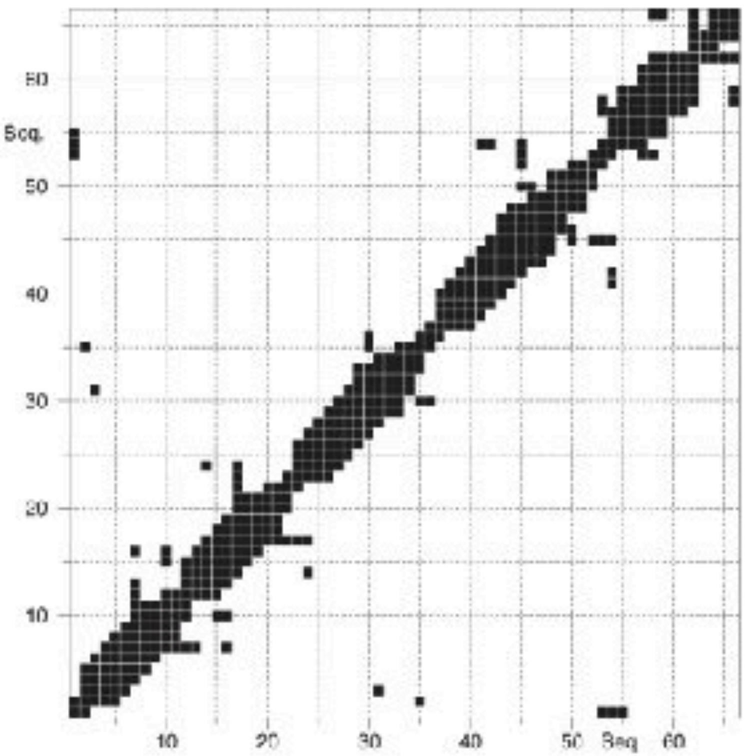




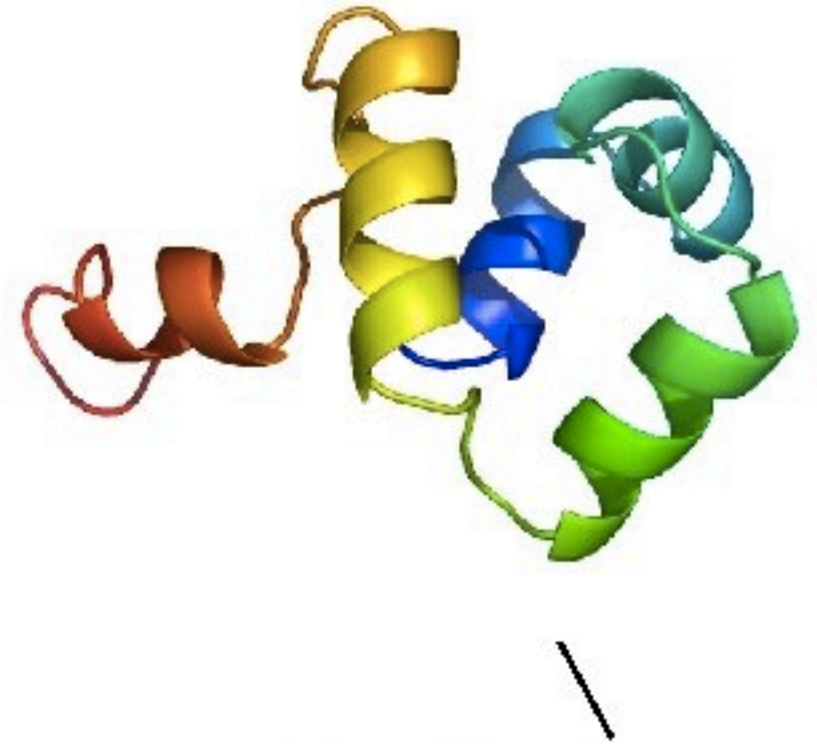
Protein contact map

In this example only contacts between residues with their Cα within 6 Å are considered

map of Cα-Cα distances < 6 Å



Both axes are the sequence of the protein



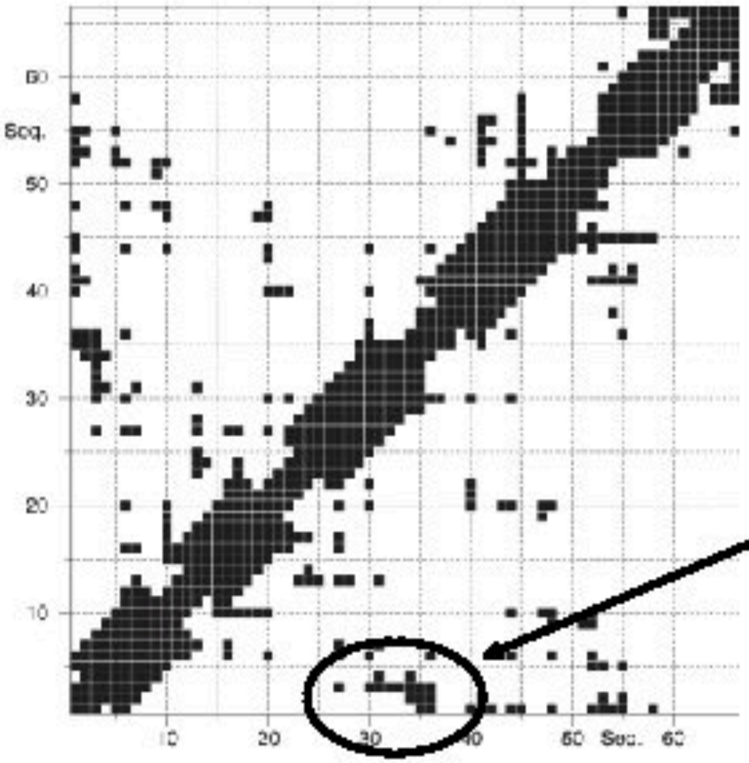
rainbow ribbon diagram
blue to red: N to C

Structure of n15 Cro

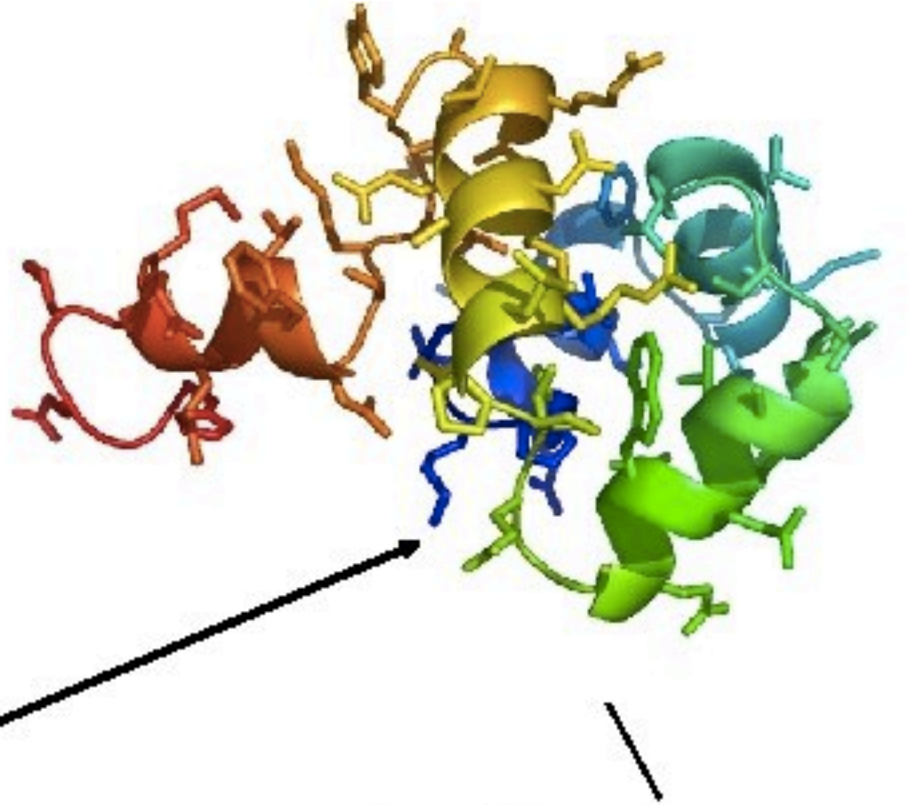
Protein contact map

In this example contacts between residues with any of their heavy (non-hydrogen) atoms within 6 Å are considered

map of all heavy atom distances < 6 Å (includes side chains)



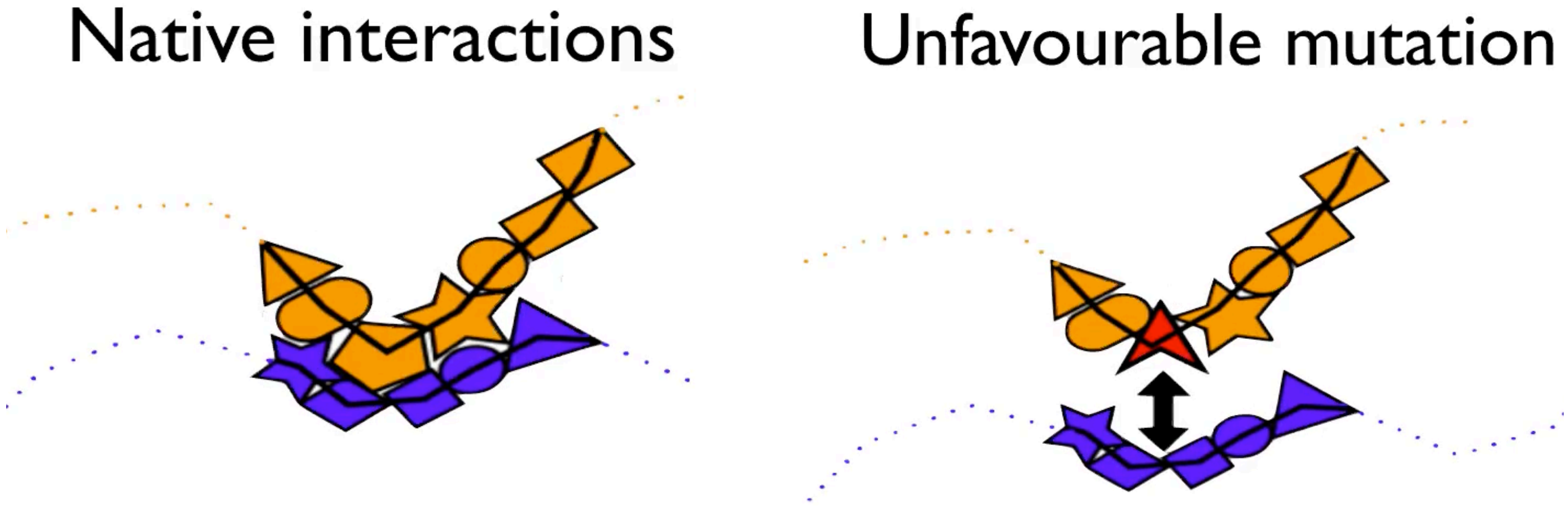
Both axes are the sequence of the protein



rainbow ribbon diagram
blue to red: N to C

Structure of n15 Cro

Residue-residue contacts

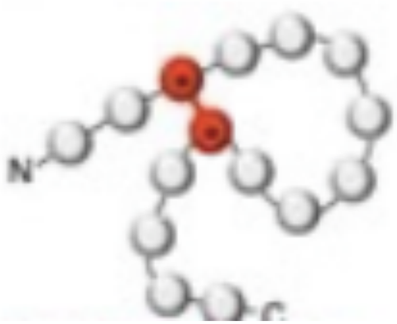


Native interactions

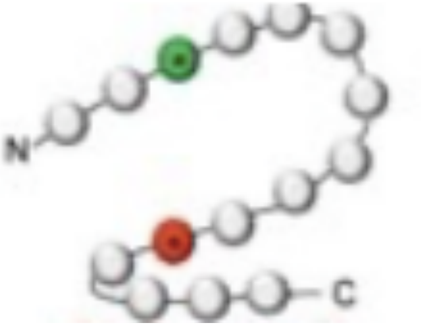
Unfavourable mutation

Residue which are in close contact tend to be complementary in shape and properties

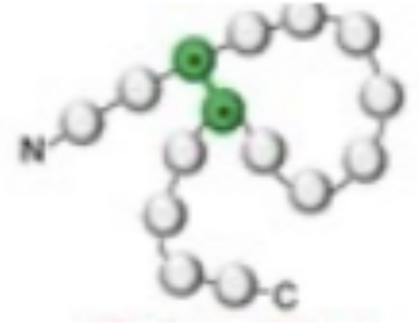
Residue-residue contacts



Initial sequence



single loss of function mutation

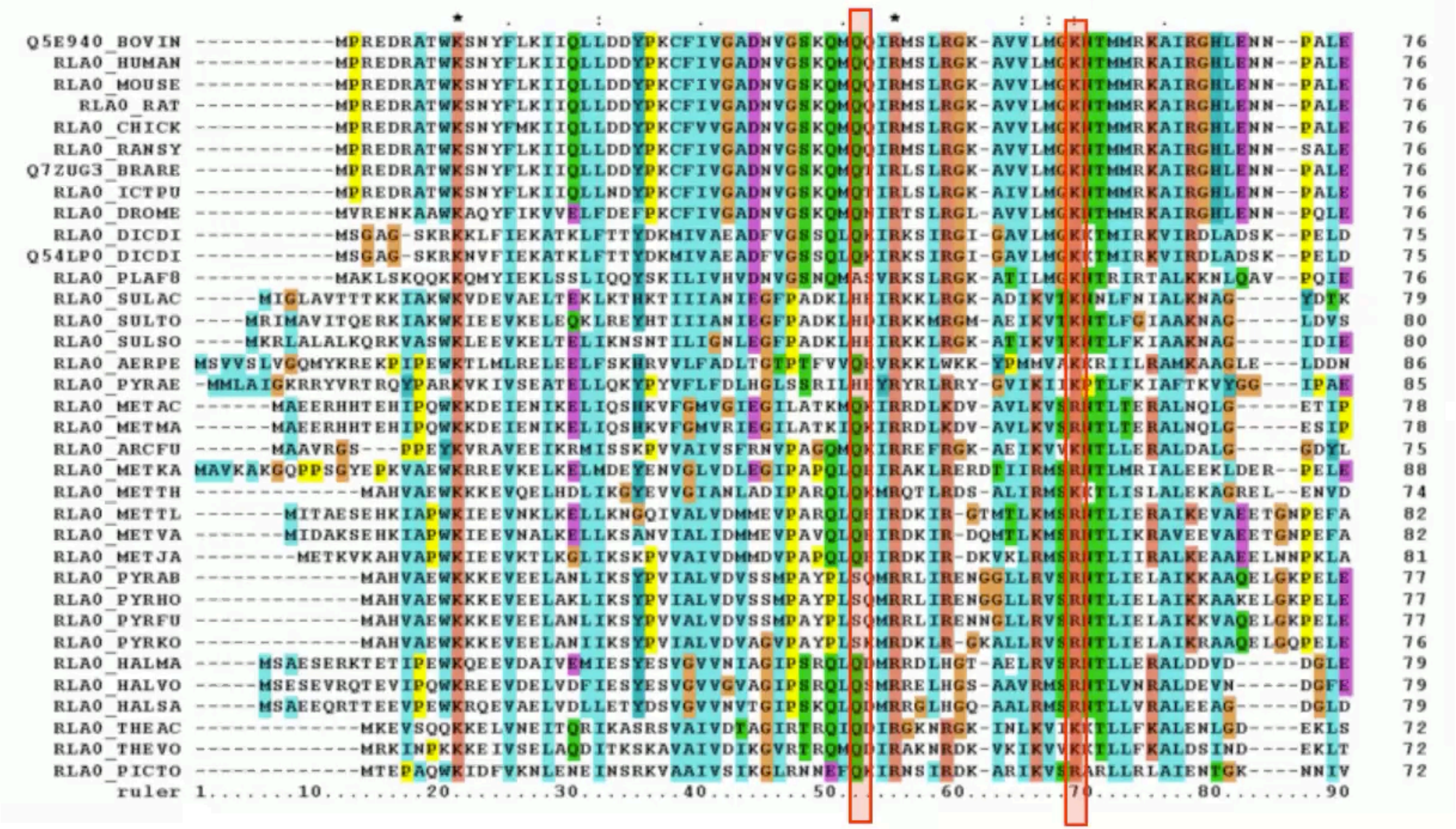


rescued by a compensating mutation

Residue which are in close contact tend to be complementary in shape and properties

If one of them gets mutated, a compensating mutation will most probably occur to the other amino acid involved in the contact

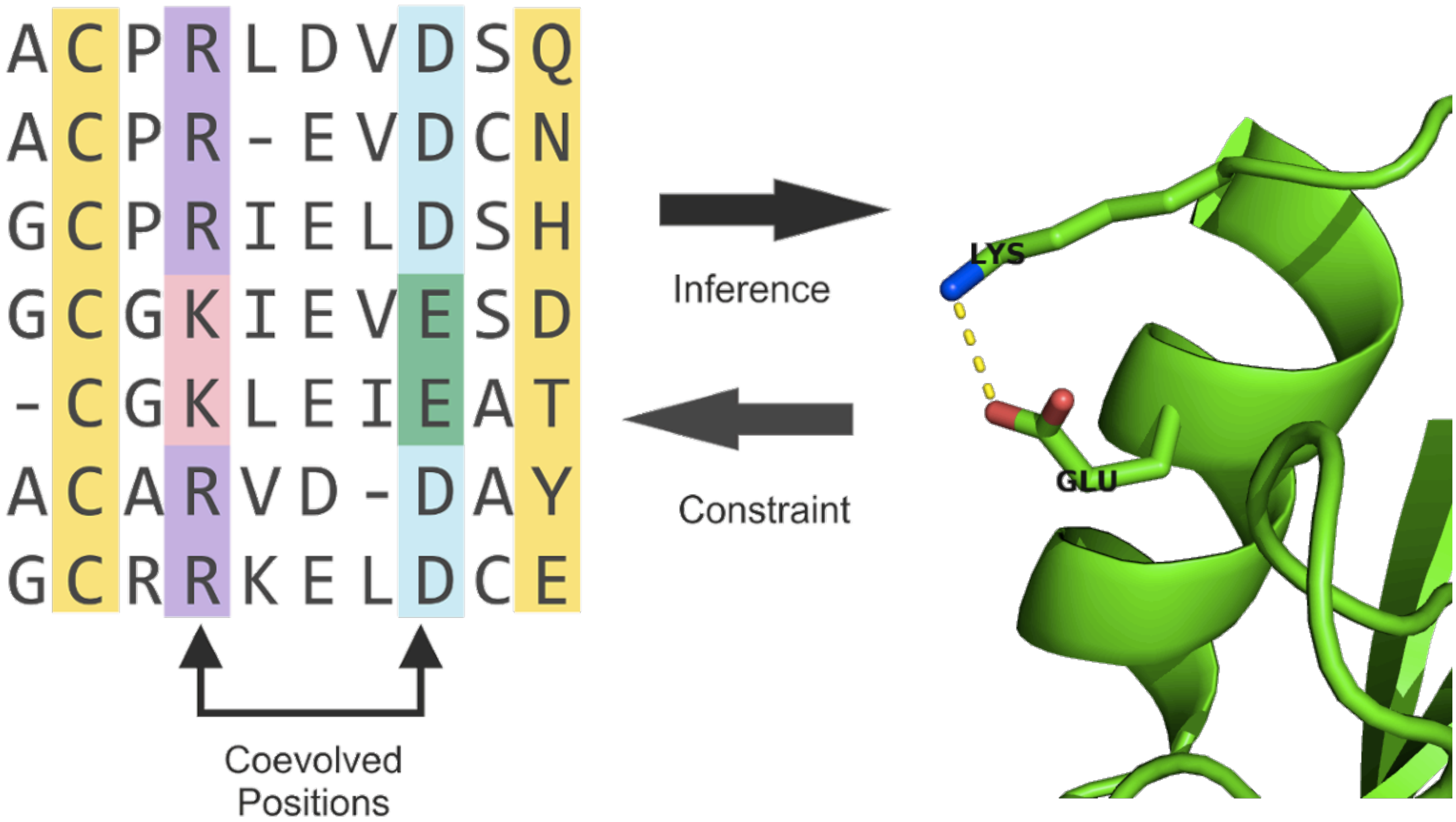
Residue-residue contact prediction



```
                                 *   .      :       .          .   *         : : .      .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE   76
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE   76
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD   75
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD   75
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE   76
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHHIRKKLRGK-ADIKVTKHNLFNIALKNAG-----YDTK   79
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKHTLFGIAAKNAG-----LDVS   80
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKHTLFKIAAKNAG-----IDIE   80
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN   86
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE   85
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRHTLTERALNQLG-----ETIP   78
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRHTLTERALNQLG-----ESIP   78
  RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKHTLLERALDALG-----GDYL   75
  RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRHTLMRIALEEKLDER--PELE   88
  RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD   74
  RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRHTLIERAIKEVAEETGNPEFA   82
  RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRHTLIKRAVEEVAEETGNPEFA   82
  RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRHTLIIRALKEAAEELNNPKLA   81
  RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRHTLIELAIKKAAQELGKPELE   77
  RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRHTLIELAIKKAAKELGKPELE   77
  RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRHTLIELAIKKVAQELGKPELE   77
  RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRHTLIELAIKRAAQELGQPELE   76
  RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRHTLLERALDDVD-----DGLE   79
  RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRHTLVNRALDEVN-----DGFE   79
  RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRHTLLVRALEEAG-----DGLD   79
  RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS   72
  RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT   72
  RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV   72
       ruler 1.......10........20........30........40........50........60........70........80........90
```

The theoretical basis for residue-residue contact prediction is that residues which are in contact tend to co-evolve, in order to stay nicely complementary

Residue-residue contact prediction



```
ACPRLDVDSQ
ACPR-EVDCN
GCPRIELDSH
GCGKIEVESD
-CGKLEIEAT
ACARVD-DAY
GCRRKELDCE
```

Inference

Constraint

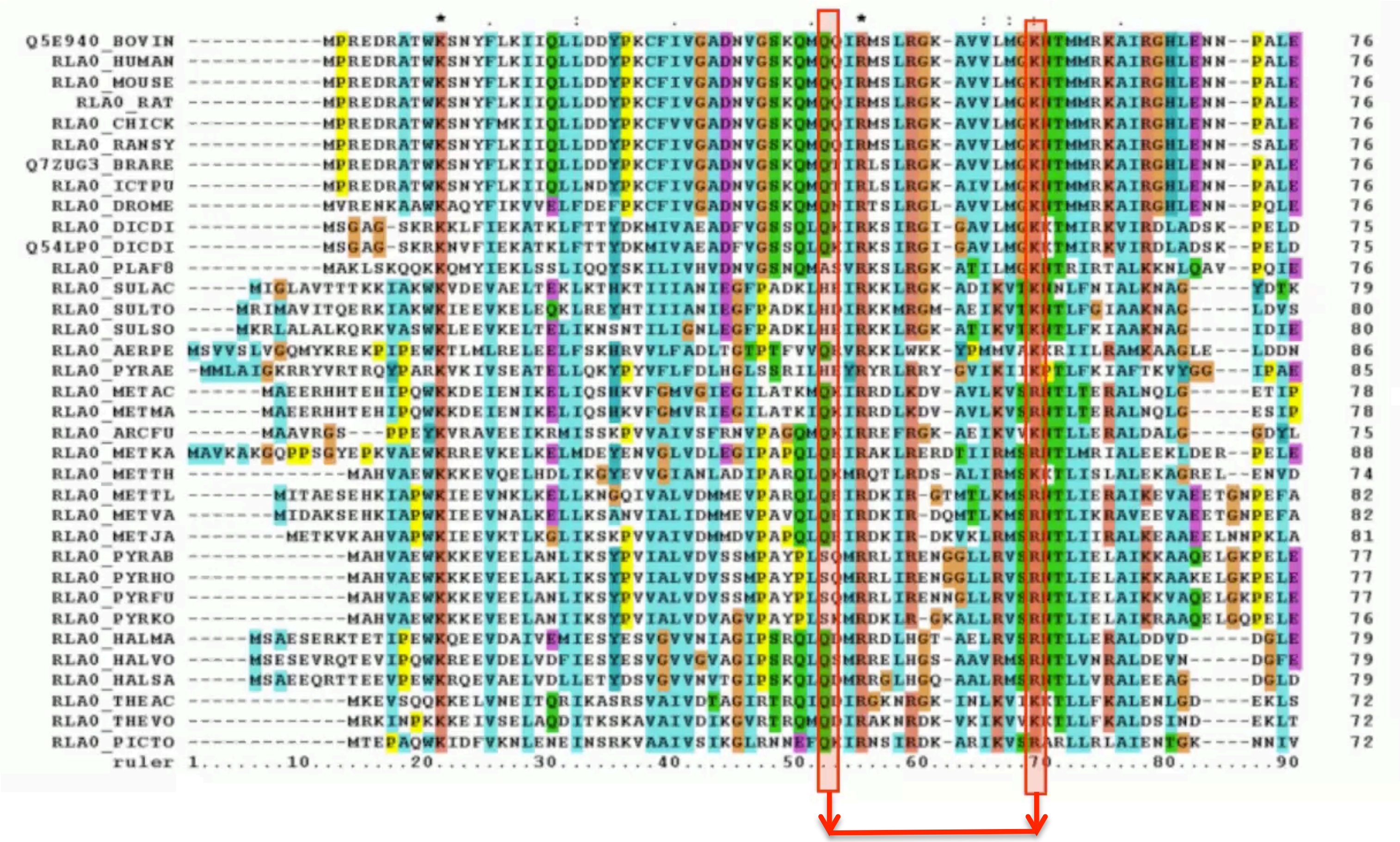
Coevolved Positions

The theoretical basis for residue-residue contact prediction is that residues which are in contact tend to co-evolve, in order to stay nicely complementary

Residue-residue contact prediction

The MSA of a protein family comprises homolog sequences from a common ancestor aligned relative to each other

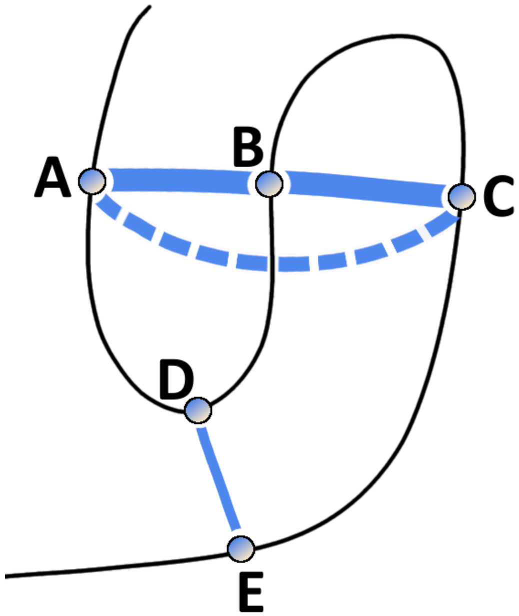
Therefore, **compensatory mutations** in MSA columns can be **used to infer spatial proximity** of residue pairs



Residue-residue contact prediction

Early contact prediction methods used local pairwise statistics to infer contacts considering pairs of amino acids as statistically independent from others

The traditional covariance approaches suffered from high false positive rates because of their inability to cope with <u>transitive effects</u> that arise from chains of correlations between multiple residue pairs



Considering three residues A, B and C, where A physically interacts with B and B with C, strong statistical dependencies between pairs (A,B) and (B,C) can induce strong <u>indirect signals for residues A and C,</u> although they are not physically interacting, which can be even larger than signals of other directly interacting pairs (D,E) and thus lead to false predictions

Residue-residue contact prediction

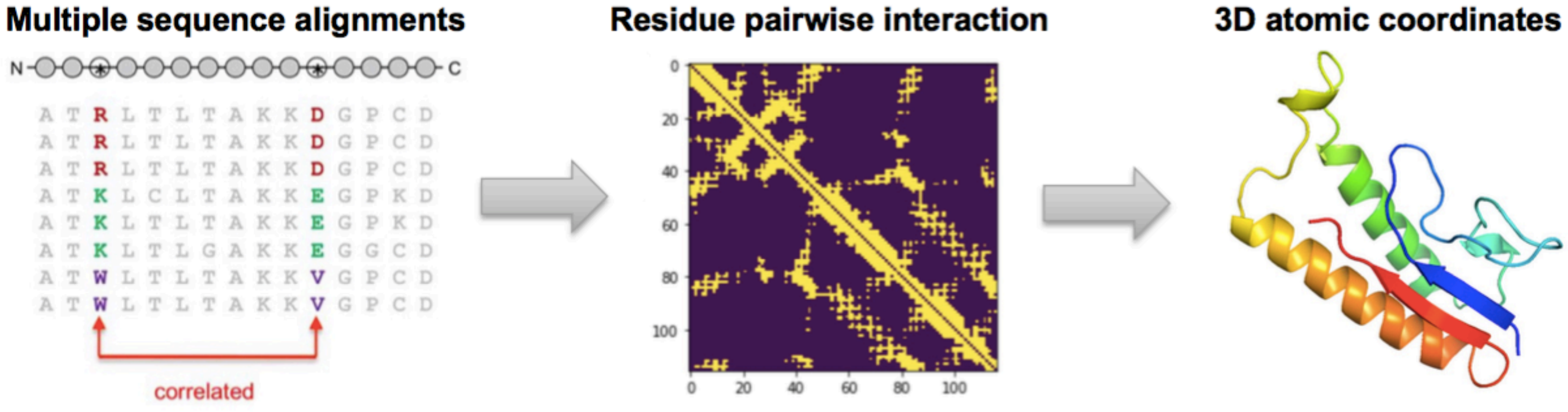
To deal with this, first a global statistical model that made predictions for a single residue pair while considering all other pairs in the protein was developed, which represented a huge leap forward

Then, machine-learning based methods, including neural networks, have emerged that extract features from MSAs in order to learn associations between input features and residue-residue contacts

Sequence features used in input typically include predicted solvent accessibility, predicted secondary structure, contact potentials, conservation scores, pairwise coevolution statistics, etc.

Residue-residue contact prediction

When residue pairwise interactions (contact maps) are predicted based on coevolution, i.e. on the MSA obtainable for a protein, they can be used for predicting its 3D structure

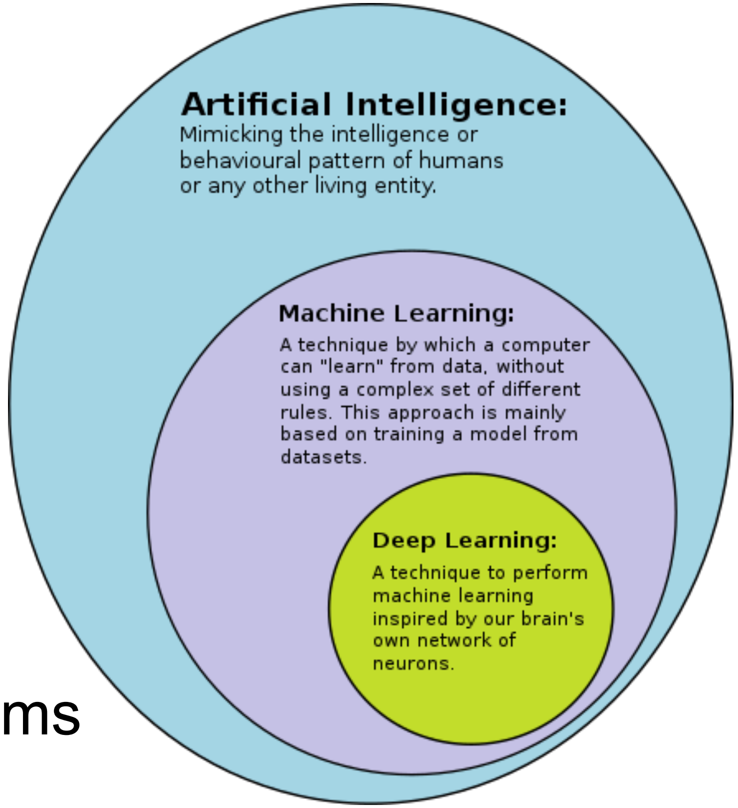

Multiple sequence alignments
N
C
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
correlated
Residue pairwise interaction
0
20
40
60
80
100
0
20
40
60
80
100
3D atomic coordinates

Deep learning

Deep learning methods are **machine learning (ML)** methods based on artificial neural networks (ANNs), also named deep neural networks (DNNs)

The adjective "deep" in deep learning refers to the use of multiple layers in the network

Since the 2010s, advances in ML algorithms and computer hardware have led to more efficient methods for training DNNs that contain many layers of non-linear hidden units


Artificial Intelligence:
Mimicking the intelligence or behavioural pattern of humans or any other living entity.
Machine Learning:
A technique by which a computer can "learn" from data, without using a complex set of different rules. This approach is mainly based on training a model from datasets.
Deep Learning:
A technique to perform machine learning inspired by our brain's own network of neurons.

Deep learning: common applications

Within science

DNNs have been successfully applied to predict the biomolecular target of a drug, to detect toxic effects of environmental chemicals in nutrients, household products and drugs, etc.

Outside science

Fraud detection
Customer relationship management systems
Computer vision
Vocal AI
Natural language processing
Autonomous vehicles
Supercomputers
Investment modeling
E-commerce

Siri, Alexa, Cortana, Google Assistant, etc., are all very popular applications of Deep Learning

Deep learning: limitations

Deep learning and neural networks in general may have two main limitations:

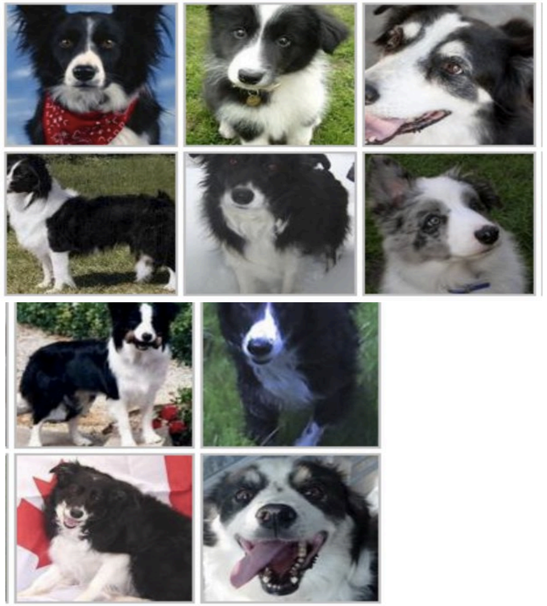
overfitting, i.e. the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit to additional data or predict future observations on unseen data; an overfitted model contains more parameters than can be justified by the data. It can be a consequence of the training data being incomplete and redundant

computational time, the more sophisticated is the network the more CPU time it will require

Data diversity (heterogenicity) *vs* overfitting

Training Data

(the most diverse the better)



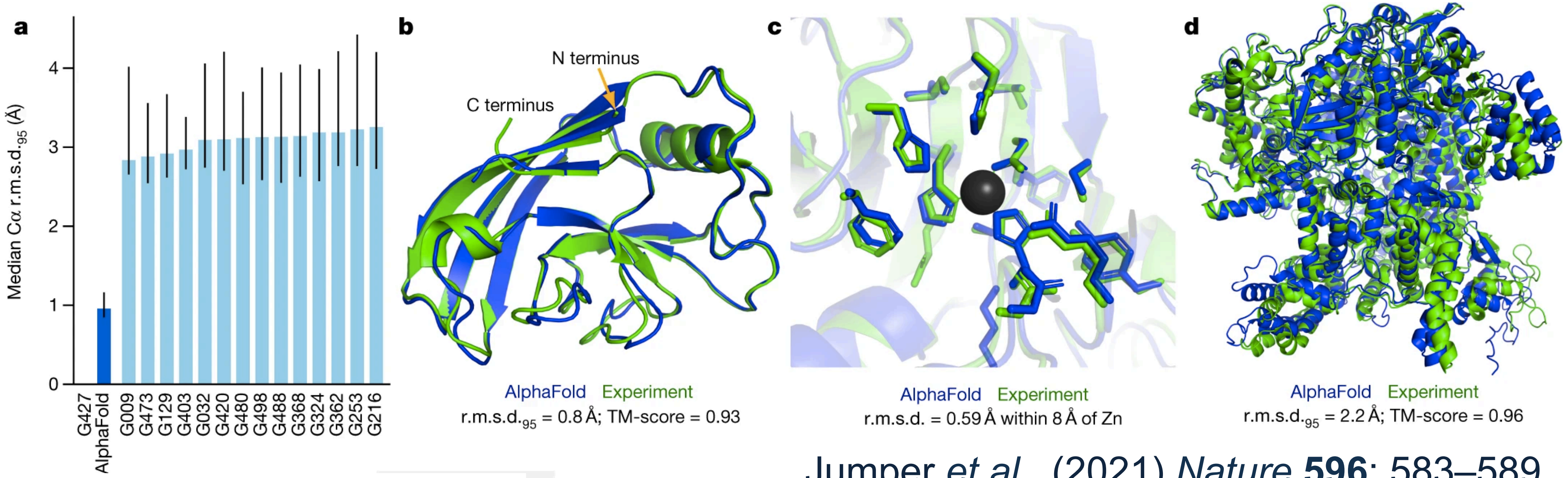
risk of overfitting: data correspond to a specific dog breed

Future observations to be predicted

Example: dog recognition

AlphaFold2: the structure prediction miracle

Performance on the CASP14 dataset (n = 87 protein domains)



AlphaFold Experiment
r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93

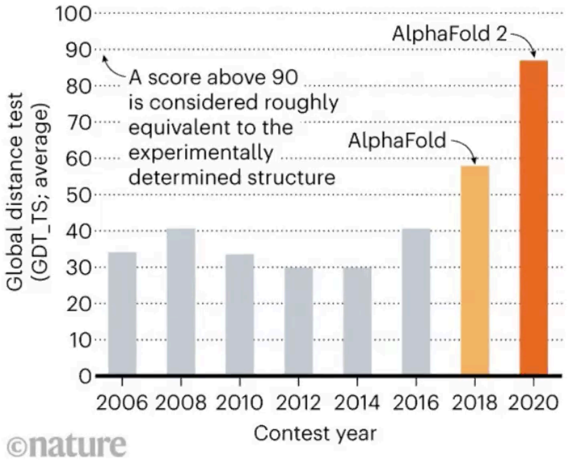
AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn

AlphaFold Experiment
r.m.s.d.$_{95}$ = 2.2 Å; TM-score = 0.96

Jumper *et al.*, (2021) *Nature* **596**: 583–589

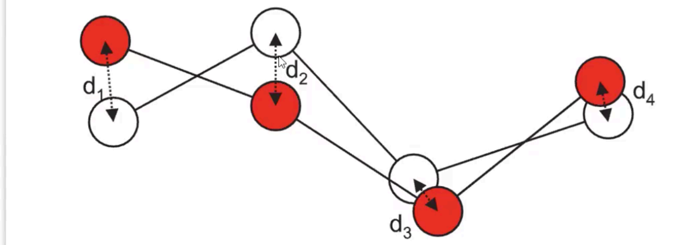
STRUCTURE SOLVER

DeepMind's AlphaFold 2 algorithm significantly outperformed other teams at the CASP14 protein-folding contest — and its previous version's performance at the last CASP.



©nature

- Root Mean Square Deviation (RMSD)



$$rmsd = \sqrt{\frac{\sum_i d_i^2}{n}}$$

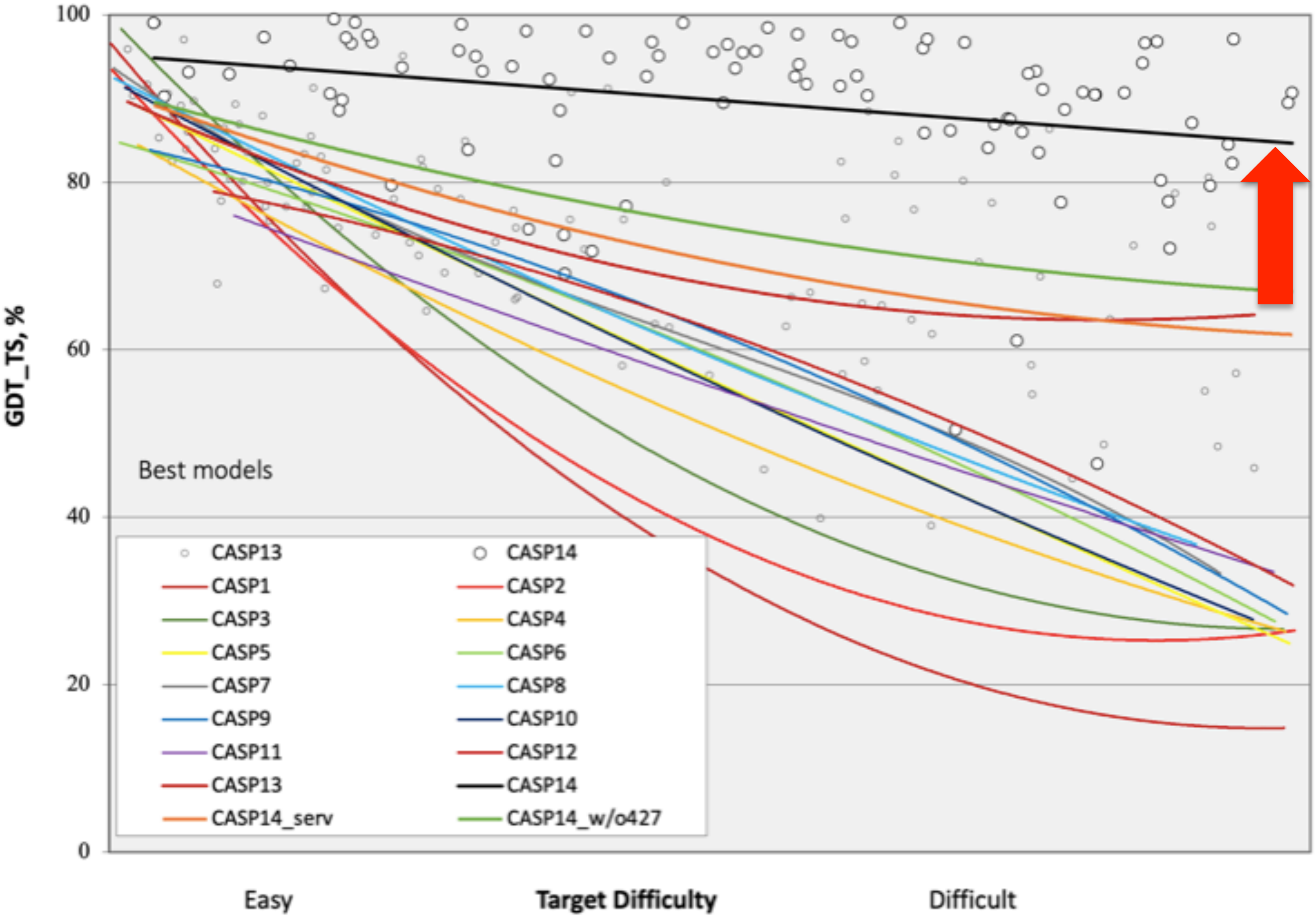
n = Number of aligned residue pairs
d = Distance between each pair of atoms

GDT_TS: percentage of corresponding α-carbons within a 4 Å distance

AlphaFold2: the structure prediction miracle

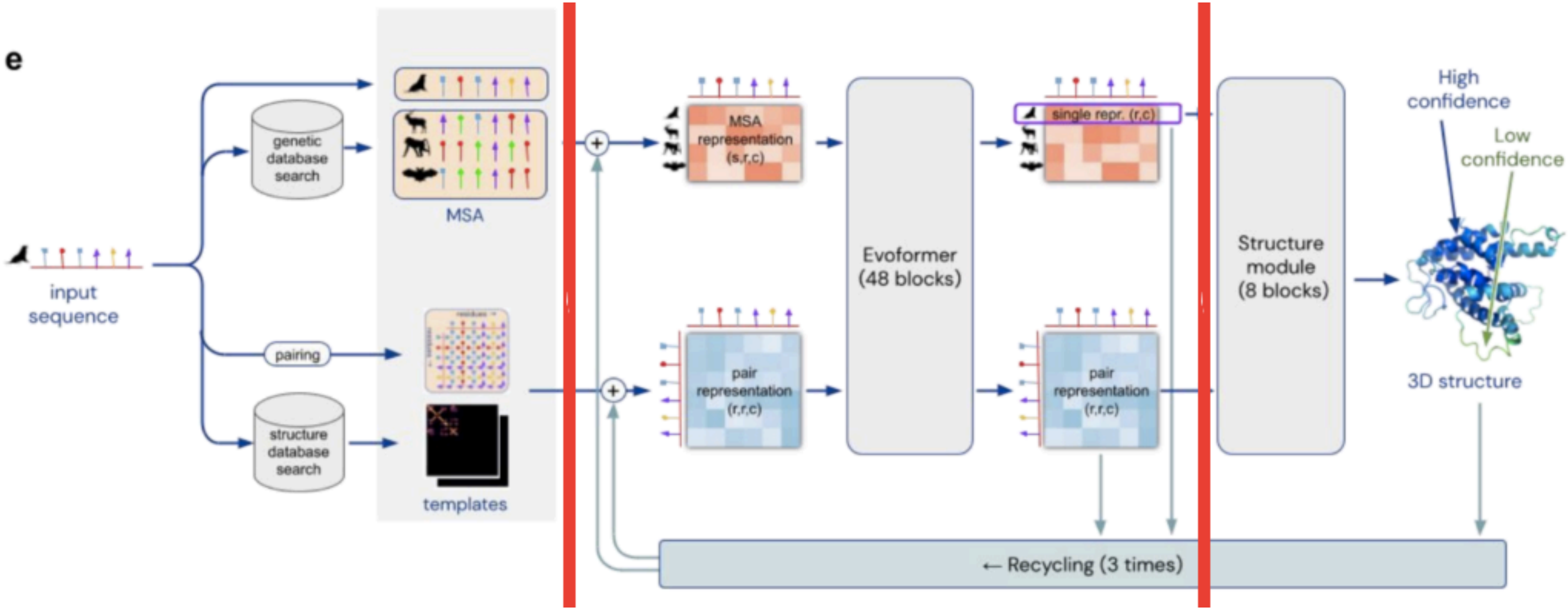
Performance on the CASP14 dataset (n = 87 protein domains)

The leap in performance in CASP14



AlphaFold2: the structure prediction miracle

overview

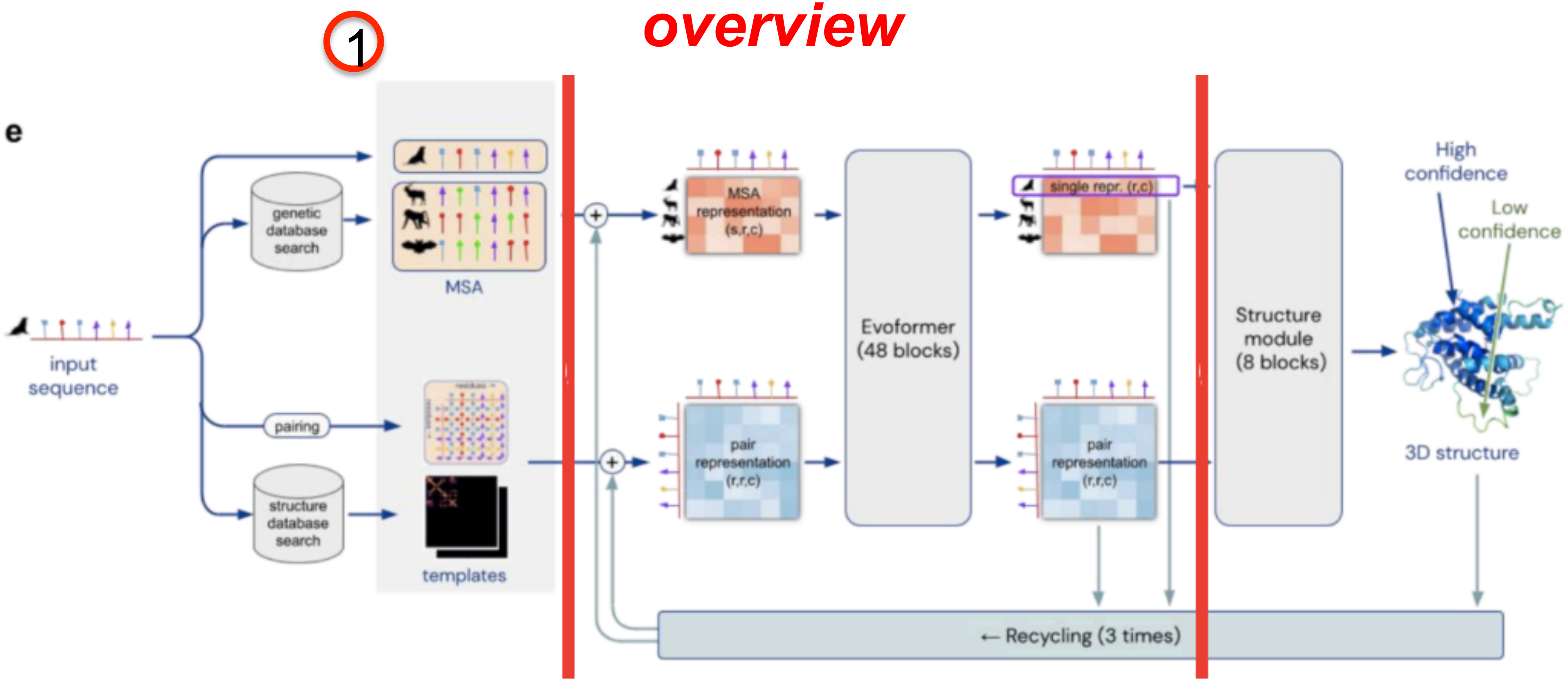


Jumper *et al.*, (2021) *Nature* **596**: 583–589

https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

AlphaFold2: the structure prediction miracle

overview

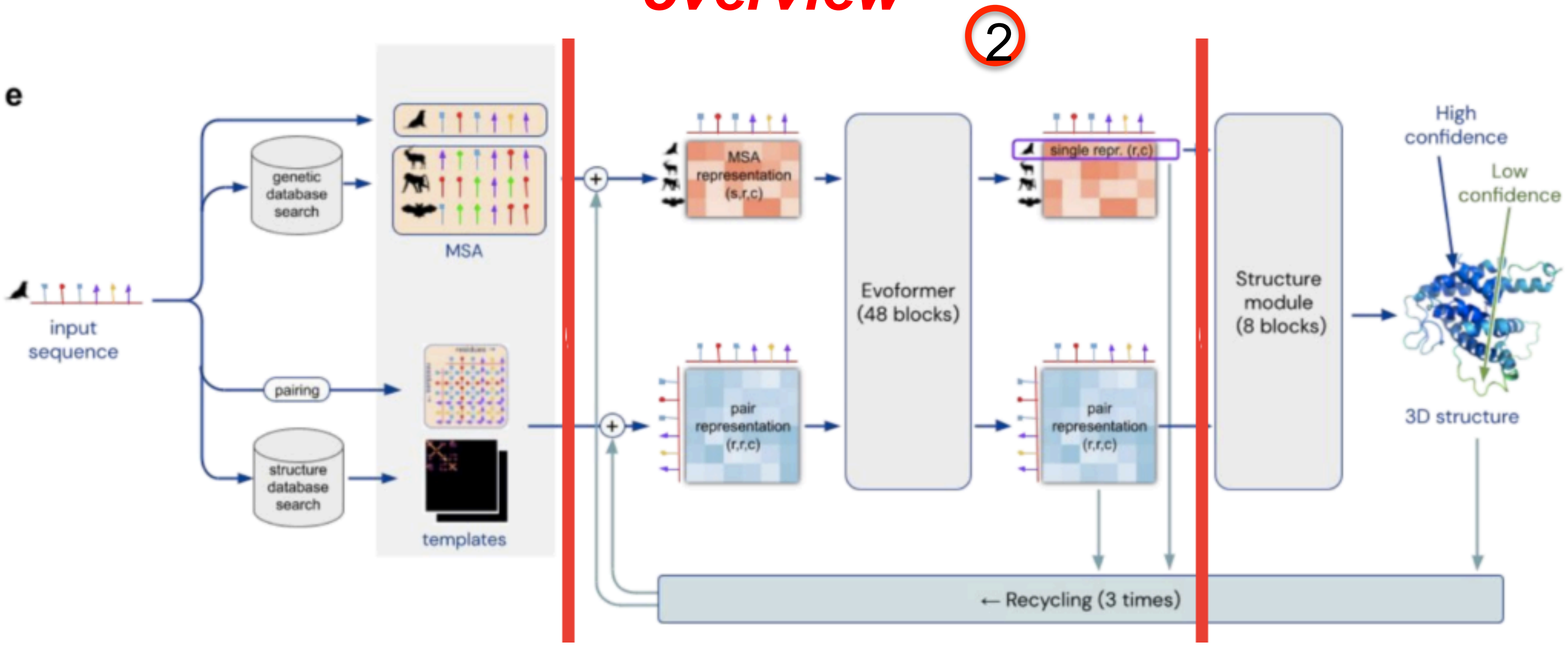


First, AlphaFold 2 uses the input amino acid sequence to query several databases of protein sequences, and constructs a multiple sequence alignment (MSA) highlighting the parts of the sequence that are more likely to mutate and possible correlations

It also tries to identify proteins that may have a similar structure to the input (“templates”), and constructs an initial representation of the structure (the “pair representation”), i.e. a model of which amino acids are likely to be in contact with each other

AlphaFold2: the structure prediction miracle

overview

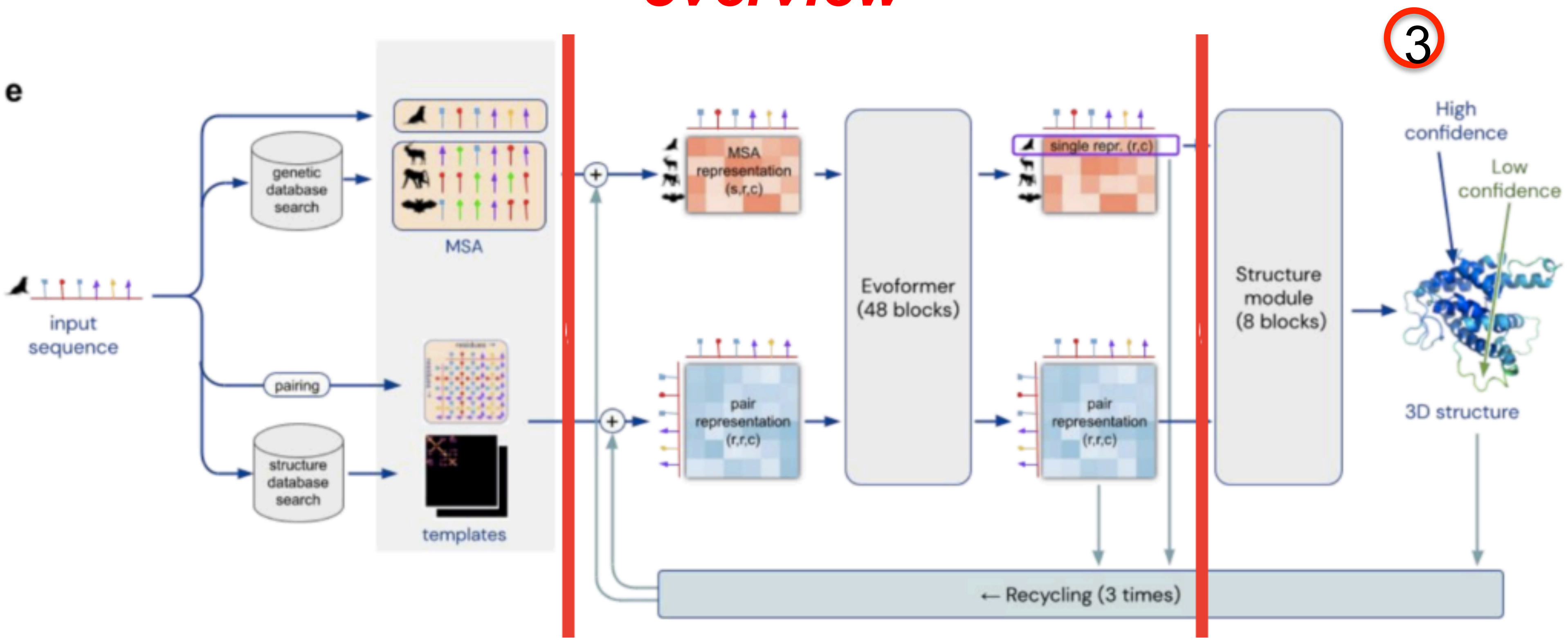


Then, AlphaFold 2 takes the MSA and the templates, and passes them through a transformer (***Evoformer, a neural network***), sort of an "oracle" that can quickly identify which pieces of information are more informative

The objective of this part is to refine the representations of the MSA and the pair interactions, and to iteratively exchange information between them. This process is organised in blocks that are repeated iteratively (48 blocks in the published model)

AlphaFold2: the structure prediction miracle

overview

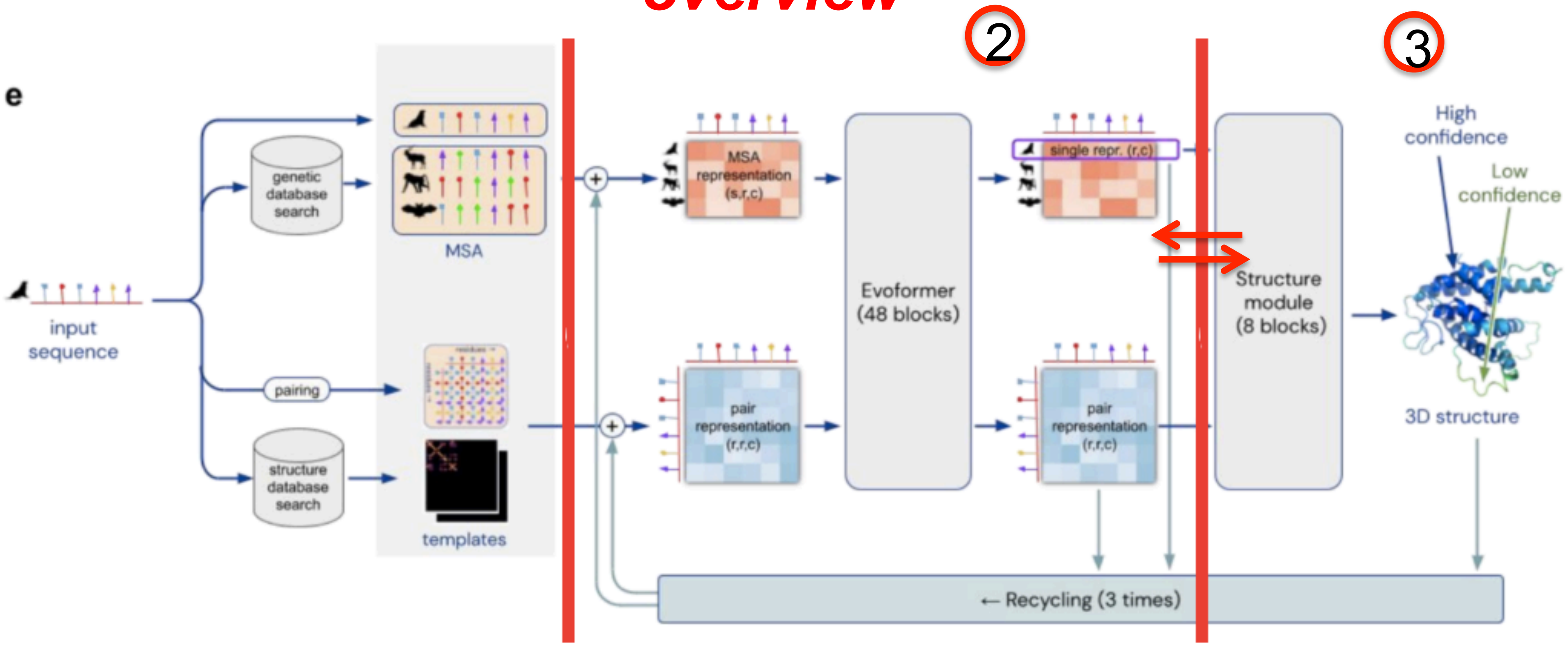


The last part is the structure module. This piece of the pipeline (again ***a neural network***) takes the refined “MSA representation” and “pair representation”, and leverages them to construct a three-dimensional model of the structure

This network does not use any optimization algorithm: it generates a final 3D structure, including side chains, in a single step

AlphaFold2: the structure prediction miracle

overview

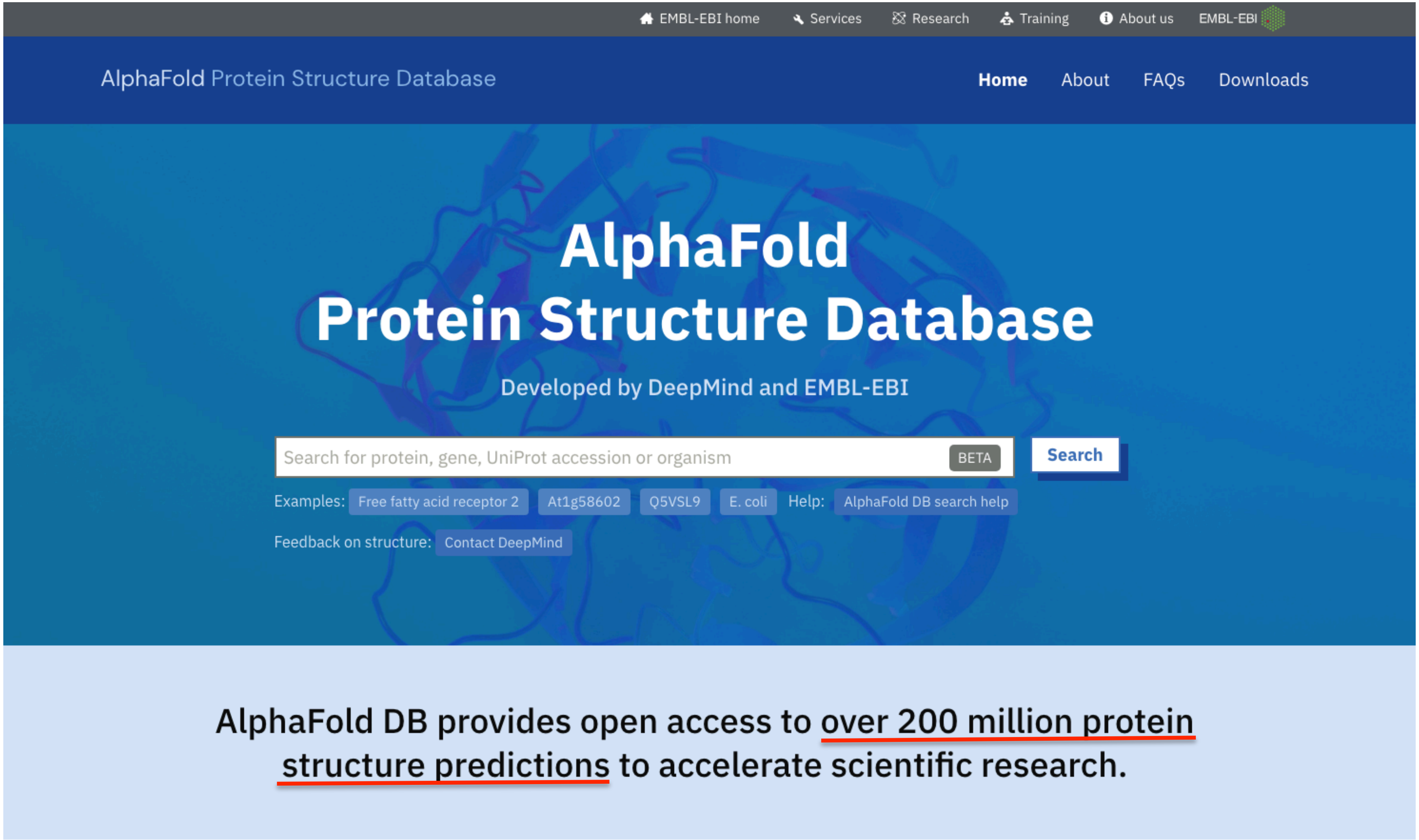


The model works iteratively. After generating a final structure, it will take all the information (i.e. MSA representation, pair representation and predicted structure) and pass it back to the beginning of the Evoformer blocks

This allows the model to refine its predictions

https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-021-03819-2/MediaObjects/41586_2021_3819_MOESM5_ESM.mp4

AlphaFold2: the DataBase



EMBL-EBI home
Services
Research
Training
About us
EMBL-EBI
AlphaFold Protein Structure Database
Home
About
FAQs
Downloads

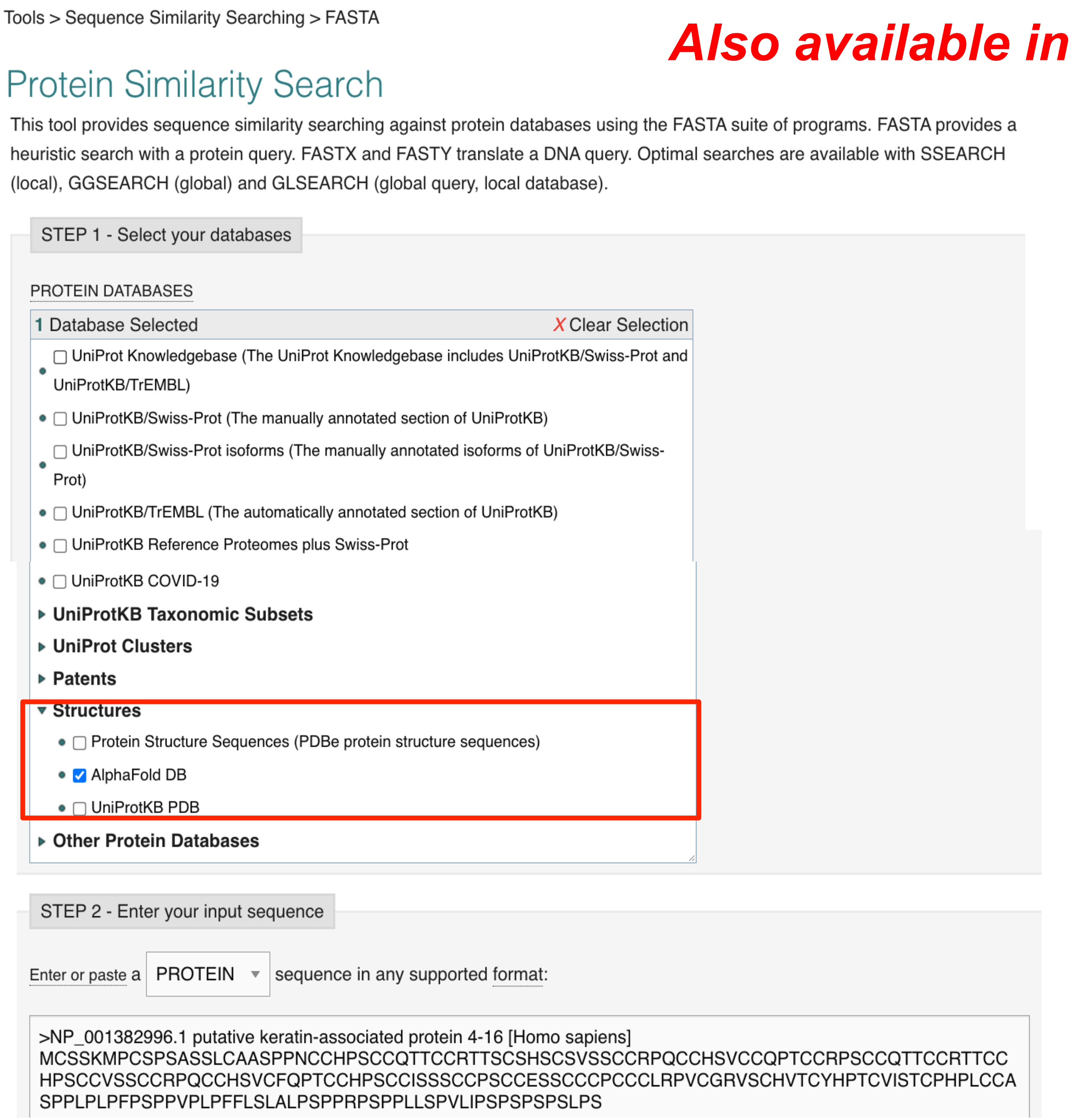
AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism
BETA
Search
Examples: Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli Help: AlphaFold DB search help
Feedback on structure: Contact DeepMind

AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.

Also available in FASTA



Tools > Sequence Similarity Searching > FASTA

Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

PROTEIN DATABASES

1 Database Selected X Clear Selection

- □ UniProt Knowledgebase (The UniProt Knowledgebase includes UniProtKB/Swiss-Prot and UniProtKB/TrEMBL)
- □ UniProtKB/Swiss-Prot (The manually annotated section of UniProtKB)
- □ UniProtKB/Swiss-Prot isoforms (The manually annotated isoforms of UniProtKB/Swiss-Prot)
- □ UniProtKB/TrEMBL (The automatically annotated section of UniProtKB)
- □ UniProtKB Reference Proteomes plus Swiss-Prot
- □ UniProtKB COVID-19

▸ **UniProtKB Taxonomic Subsets**

▸ **UniProt Clusters**

▸ **Patents**

▾ **Structures**

- □ Protein Structure Sequences (PDBe protein structure sequences)
- ☑ AlphaFold DB
- □ UniProtKB PDB

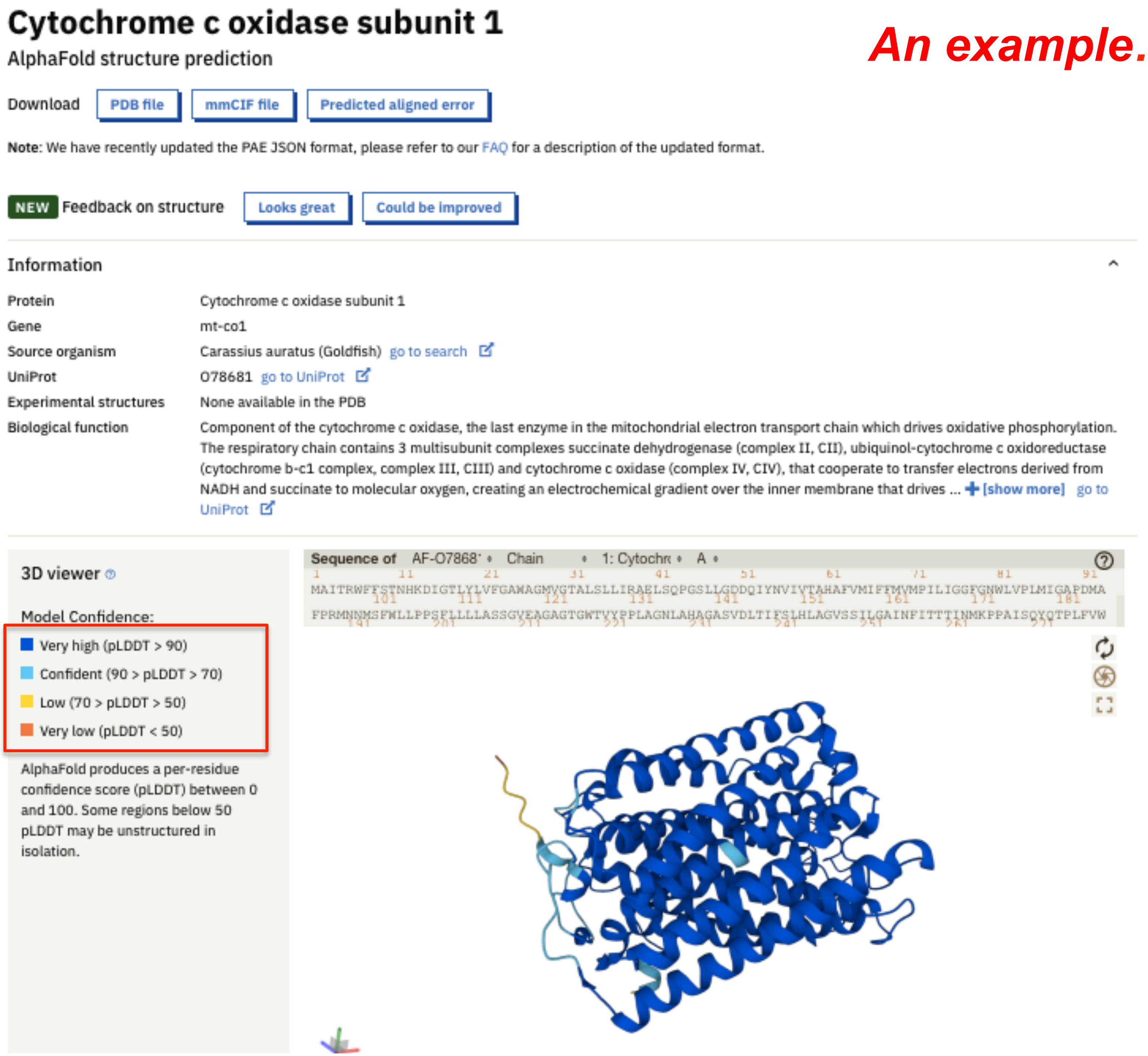
▸ **Other Protein Databases**

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN ▾ sequence in any supported format:

```
>NP_001382996.1 putative keratin-associated protein 4-16 [Homo sapiens]
MCSSKMPCSPSASSLCAASPPNCCHPSCCQTTCCRTTSCSHSCSVSSCCRPQCCHSVCCQPTCCRPSCCQTTCCRTTCC
HPSCCVSSCCRPQCCHSVCFQPTCCHPSCCISSSCCPSCCESSCCCPCCCLRPVCGRVSCHVTCYHPTCVISTCPHPLCCA
SPPLPLPFPSPPVPLPFFLSLALPSPPRPSPPLLSPVLIPSPSPSPSLPS
```

An example…



Cytochrome c oxidase subunit 1

AlphaFold structure prediction

Download PDB file mmCIF file Predicted aligned error

Note: We have recently updated the PAE JSON format, please refer to our FAQ for a description of the updated format.

NEW Feedback on structure Looks great Could be improved

Information

Protein	Cytochrome c oxidase subunit 1
Gene	mt-co1
Source organism	Carassius auratus (Goldfish) go to search
UniProt	O78681 go to UniProt
Experimental structures	None available in the PDB
Biological function	Component of the cytochrome c oxidase, the last enzyme in the mitochondrial electron transport chain which drives oxidative phosphorylation. The respiratory chain contains 3 multisubunit complexes succinate dehydrogenase (complex II, CII), ubiquinol-cytochrome c oxidoreductase (cytochrome b-c1 complex, complex III, CIII) and cytochrome c oxidase (complex IV, CIV), that cooperate to transfer electrons derived from NADH and succinate to molecular oxygen, creating an electrochemical gradient over the inner membrane that drives ... + [show more] go to UniProt

3D viewer

Model Confidence:

- 🟦 Very high (pLDDT > 90)
- 🟦 Confident (90 > pLDDT > 70)
- 🟨 Low (70 > pLDDT > 50)
- 🟧 Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Sequence of AF-O7868 Chain 1: Cytochr A

```
1        11        21        31        41        51        61        71        81        91
MAITRWFFSTNHKDIGTLYLVFGAWAGMVGTALSLLIRAELSQPGSLLGDDQIYNVIVTAHAFVMIFFMVMPILIGGFGNWLVPLMIGAPDMA
     101       111       121       131       141       151       161       171       181
FPRMNNMSFWLLPPSFLLLLASSGVEAGAGTGWTVYPPLAGNLAHAGASVDLTIFSLHLAGVSSILGAINFITTTINMKPPAISQYQTPLFVW
   191       201       211       221       231       241       251       261       271
```

AlphaFold2 confidence score

The AlphaFold2 confidence score is the **pLDDT: predicted Local-Distance Difference Test**

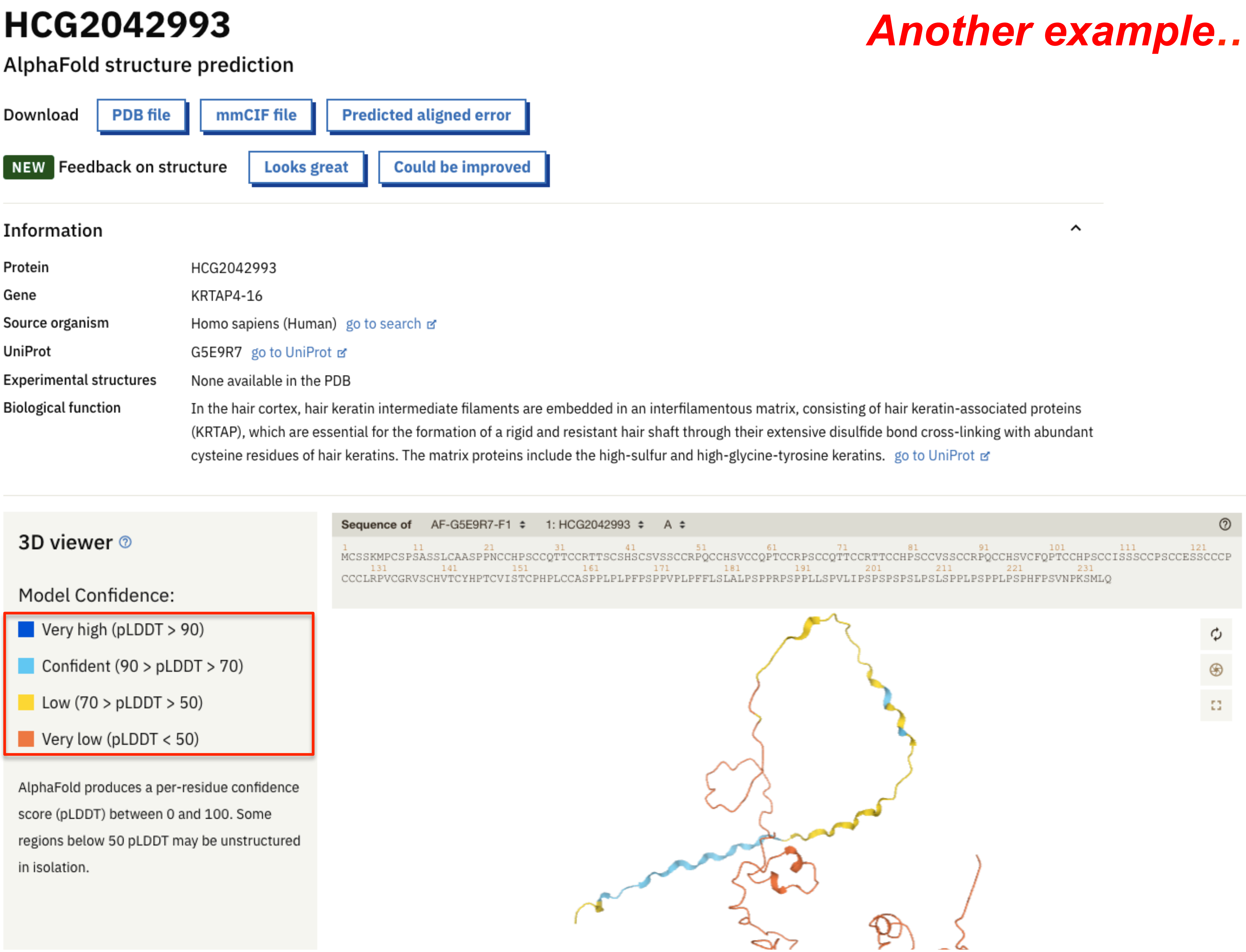
- Regions with **pLDDT > 90** are expected to be modelled to high accuracy. These should be suitable for any application that benefits from high accuracy (e.g. characterizing binding sites)

- Regions with **pLDDT between 70 and 90** are expected to be modelled well (a generally good backbone prediction)

- Regions with **pLDDT between 50 and 70** are low confidence and should be treated with caution.

- Regions with **pLDDT < 50** should not be considered, they are most probably unstructured (disordered) in physiological conditions or only structured as part of a complex

Another example…



HCG2042993

AlphaFold structure prediction

Download **PDB file** **mmCIF file** **Predicted aligned error**

NEW Feedback on structure **Looks great** **Could be improved**

Information

Protein	HCG2042993
Gene	KRTAP4-16
Source organism	Homo sapiens (Human) go to search
UniProt	G5E9R7 go to UniProt
Experimental structures	None available in the PDB
Biological function	In the hair cortex, hair keratin intermediate filaments are embedded in an interfilamentous matrix, consisting of hair keratin-associated proteins (KRTAP), which are essential for the formation of a rigid and resistant hair shaft through their extensive disulfide bond cross-linking with abundant cysteine residues of hair keratins. The matrix proteins include the high-sulfur and high-glycine-tyrosine keratins. go to UniProt

3D viewer

Model Confidence:

- 🟦 Very high (pLDDT > 90)
- 🟦 Confident (90 > pLDDT > 70)
- 🟨 Low (70 > pLDDT > 50)
- 🟧 Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Sequence of AF-G5E9R7-F1 ⇕ 1: HCG2042993 ⇕ A ⇕

```
1         11        21        31        41        51        61        71        81        91        101       111       121
MCSSKMPCSPSASSLCAASPPNCCHPSCCQTTCCRTTSCSHSCSVSSCCRPQCCHSVCCQPTCCRPSCCQTTCCRTTCCHPSCCVSSCCRPQCCHSVCFQPTCCHPSCCISSSCCPSCCESSCCCP
    131       141       151       161       171       181       191       201       211       221       231
CCCLRPVCGRVSCHVTCYHPTCVISTCPHPLCCASPPLPLPFFPSPPVPLPFFLSLALPSPPRPSPPLLSPVLIPSPSPSPSLPSLSPPLPSPPLPSPHFPSVNPKSMLQ
```

Lesson 12. Content

1. Neural networks (NNs). Mimic physiological NNs. Part of Artificial Intelligence (AI) methods. Can learn from their own errors. Need many diverse examples with a known answer to learn (to be trained) from; when complex (multilayers *etc.*) need high computational power
2. Prediction of secondary structure. Highly efficient. Performed based on NNs since at least two decades.
3. Protein contact prediction. Recently recognized as an efficient basis for protein 3D structure prediction. Exploits evolutionary info through co-evolution in MSAs.
4. 3D structure prediction with Deep Learning. Come into field in the last few years, set a revolution in it. Reaches experimental-like accuracy in most cases.