(Aspekte der Thermodynamik in der Strukturbiologie)

Einführung in die Bioinformatik

Wintersemester 2012/13

Peter Güntert

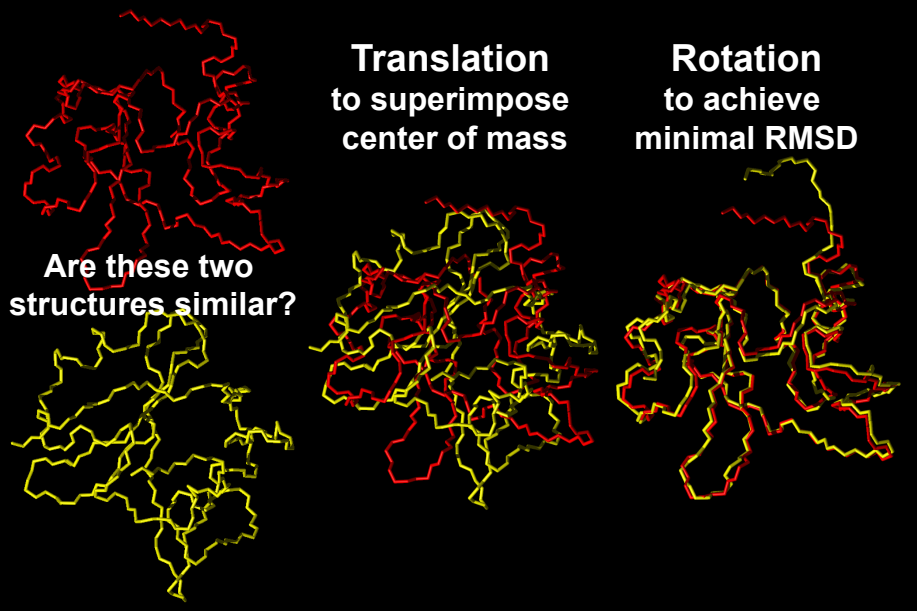
Protein Structure Similarity

Outline

- Structure comparison
- Structural similarity search

Structure comparison

Optimal superposition of structures



Measures of structural similarity

- **RMSD:** Average (root-mean-square) deviation of atom positions
- **GDT-TS:** Percentage of residues that can be superimposed under given distance cutoffs

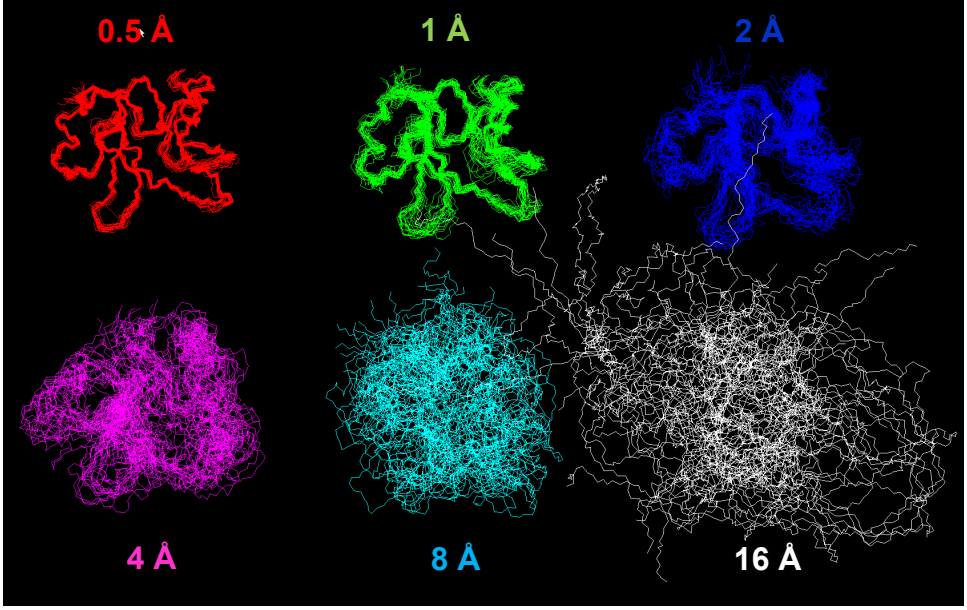
RMSD (root-mean-square deviation)

- Zwei Strukturen mit *n* Atomen und Koordinaten $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ und $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$

$$RMSD = \min_{R,\vec{t}} \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\vec{x}_i - R\vec{y}_i - \vec{t}\right|^2}$$

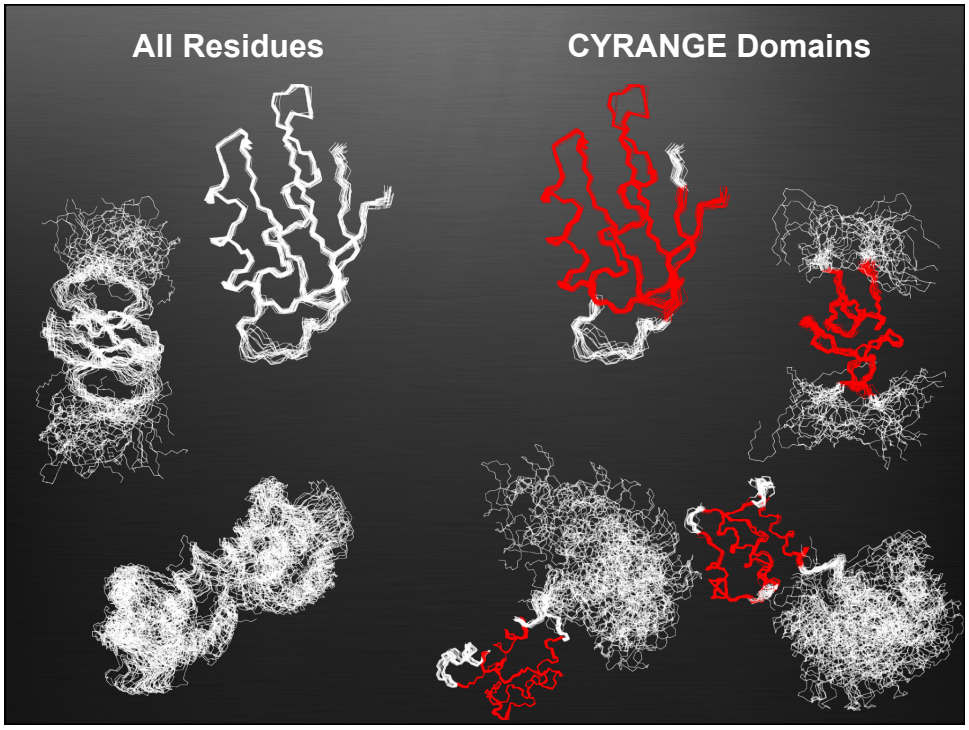
- Minimum über alle Rotationen *R* und Translationen $\boldsymbol{t}$ → optimale Überlagerung

RMSD values of structure bundles



Residue ranges for superpositions

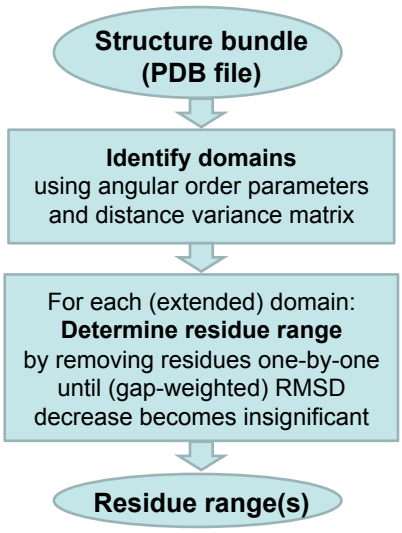
- RMSD values and superposition results are dominated by the most divergent parts of the structures.
- Unstructured parts of a protein (e.g. flexible tails, disordered loops, surface side-chains) should be excluded from RMSD calculation.
- Selecting proper residue ranges is important for meaningful RMSD calculations and structure superpositions.
- RMSD are given for a particular residue range and only for backbone atoms, or all "heavy" (non-hydrogen) atoms.



Residue ranges for superpositions

- **Important for RMSD calculations, structure superpositions, structure validation**
- **Inconsistent choices of residue ranges → not comparable validation results**
- **Angular order parameters not suitable because they measure local order only**
- **Automated residue range determination:**
  - without protein-specific parameter adjustment
  - ranges as large and as simple as possible
  - find domains in multi-domain proteins
  - for high- and low-resolution structure bundles

CYRANGE algorithm for residue range determination



NMRCORE/NMRCLUST
Kelley, Gardner, Sutcliffe
Protein Eng. 9, 1063 (1996)
Protein Eng. 10, 737 (1997)

Donata Kirchner
BMC Bioinformatics
12, 170 (2011)

Cluster Selection

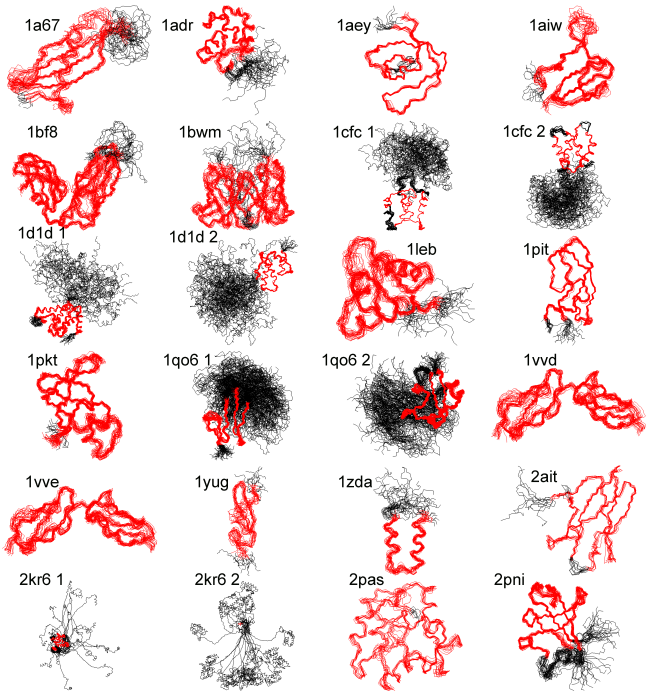
A B C D E

A B C D E

A B E C D

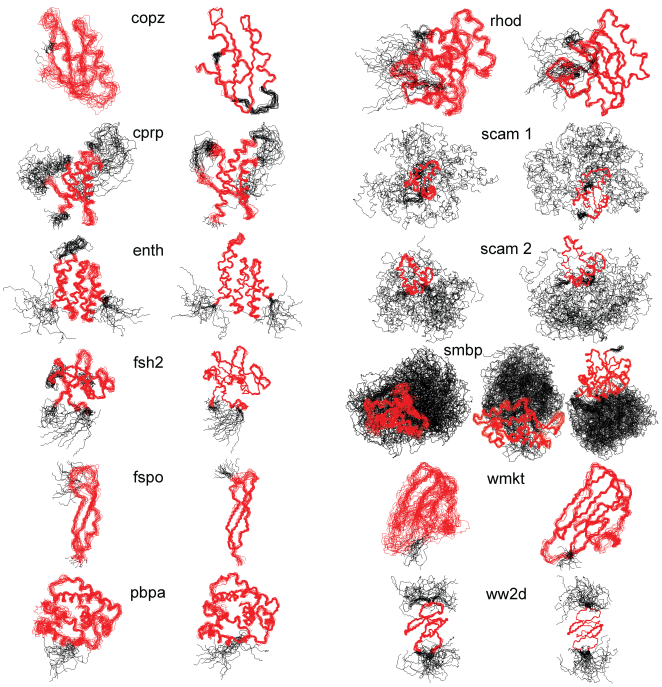
A B E C D

A C B D E



RMSD Ranges

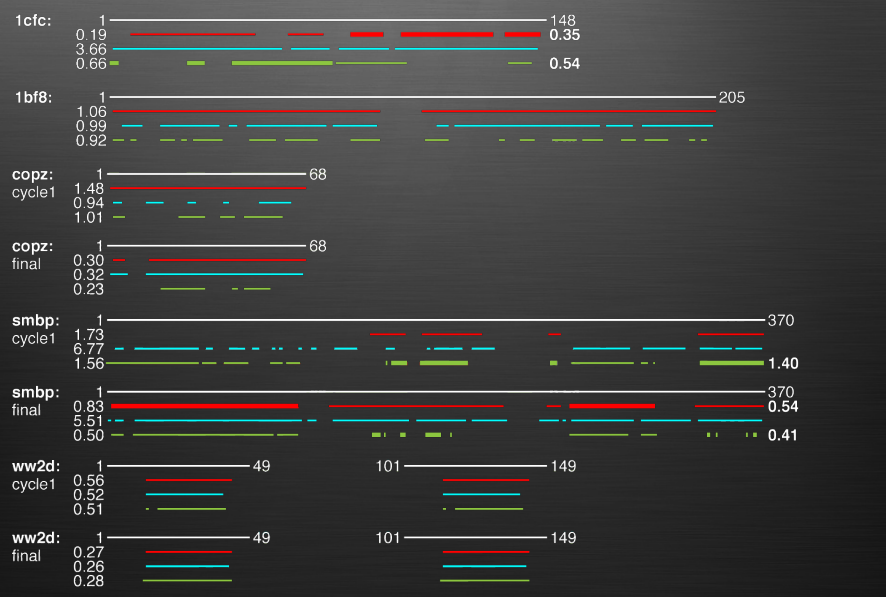
Donata Kirchner
BMC Bioinformatics
12, 170 (2011)



RMSD Ranges

Application to low-resolution (cycle 1) and high-resolution (final) structures.

Donata Kirchner

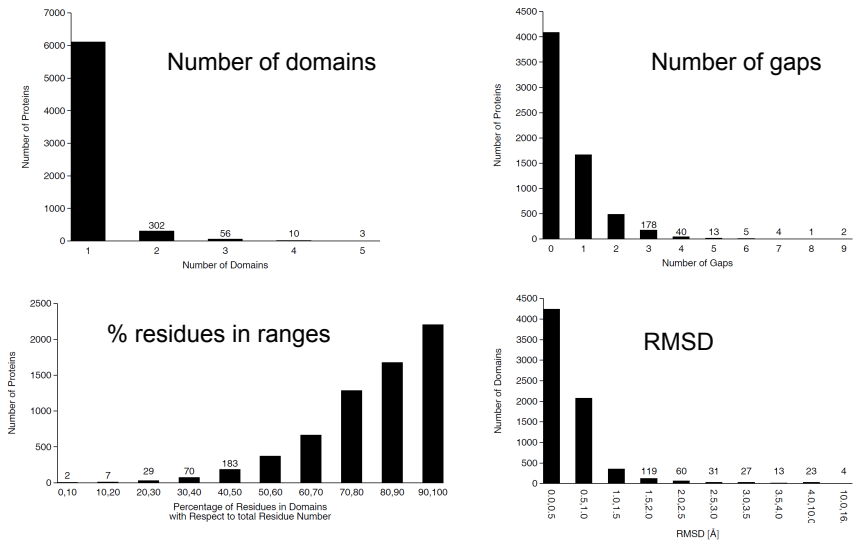


[1] Bhattacharya A, Tejero R, Montelione GT, *Proteins* 2007, 66:778–795.
[2] Snyder DA, Montelione GT, *Proteins* 2005, 59:673–686.

	Average Sequence Coverage[1]	Average RMSD[1]
CYRANGE	85 %	0.77 Å
PSVS	67 %	1.72 Å
FindCore	58 %	0.73 Å

[1] of 37 protein structures

CYRANGE residue ranges for 6483 NMR structure bundles in the PDB



Protein Domain Identification with CYRANGE

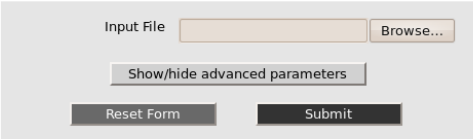
CYRANGE is used for identification of domains in an NMR-derived protein structure bundle. These domains comprise appropriate residue ranges for RMSD calculation.

If you use CYRANGE for your research projects we kindly ask you to cite the following publication:

Input Section

Providing an input file name is compulsory. All other fields contain the respective parameter's default value, which will be employed unless a different value was input. The meanings of the various parameters are explained if you move the mouse over the parameter name in question. Alternatively you will find the explanations here.

Press Submit when you are ready. Once the results have been computed you will be redirected to another page.



www.bpc.uni-frankfurt.de/cyrange.html

CYRANGE Results for smbp_final.pdb*

The following 2 domains were identified:

Residue Range	RMSD [Å]	Gaps	Number of Residues
4-109, 261-309	0.54	1	155
126-224, 248-256, 331-370	0.83	2	148

* gap: 0.4, decrease: 1.2, absdecrease: 1.6, gapwidth: 3, buffer: 3, minimum cluster size: 8

Alternative measure for structure similarity

GDT_TS

- The GDT ("global distance test") algorithm searches for the largest (not necessarily continuous) set of residues that deviate by no more than a specified distance cutoff.
- Results are reported as the percentage of residues under a given distance cutoff.
- A popular measure is the "GDT total score",

$$GDT\_TS = (P_1 + P_2 + P_4 + P_8)/4,$$

where P_d is the fraction of residues that can be superimposed under a distance cutoff of d Å, which reduces the dependence on the choice of the cutoff by averaging over four different distance cutoff values.

Structural similarity search

DALI: structure similarity search

With a rapidly growing pool of known tertiary structures, the importance of protein structure comparison parallels that of sequence alignment.

DALI algorithm for optimal pairwise alignment of protein structures:
(L. Holm & C. Sander: Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233, 123-138 (1993)):

- Coordinates of each protein are used to calculate residue-residue (C^{α}-C^{α}) distance matrices.
- Distance matrices are first decomposed into elementary contact patterns, e.g. hexapeptide-hexapeptide submatrices.
- Then, similar contact patterns in the two matrices are paired and combined into larger consistent sets of pairs. A Monte Carlo procedure is used to optimize a similarity score. Several alignments are optimized in parallel, leading to simultaneous detection of the best, second-best and so on solutions.
- DALI allows sequence gaps of any length, reversal of chain direction and free topological connectivity of aligned segments. Sequential connectivity can be imposed as an option.
- DALI is fully automatic and identifies structural resemblances and common structural cores accurately and sensitively.

DALI: structure similarity search

Dali server

Institute of Biotechnology

SERVICES & TOOLS | GROUP MEMBERS | NEWS & VACANCIES | RESEARCH | PUBLICATIONS

Protein Structure Database Searching by DaliLite v. 3

The Dali server is a network service for comparing protein structures in 3D. You submit the coordinates of a query protein structure and Dali compares them against those in the Protein Data Bank (PDB). You receive an email notification when the search has finished. In favourable cases, comparing 3D structures may reveal biologically interesting similarities that are not detectable by comparing sequences.

Requests can also be submitted by e-mail to *dali-server at helsinki dot fi*. The body of the e-mail message must contain atomic coordinates in PDB format.

If you want to know the structural neighbours of a protein already in the Protein Data Bank (PDB), you can find them in the Dali Database.

If you want to superimpose two particular structures, you can do it in the pairwise DaliLite server.

Upload a structure:

Browse···

Or enter PDB identifier: **chain:** (optional)

(Keyword search for PDB identifiers)

Job name:

(optional)

Enter email address for notification:

(recommended)

submit clear

http://ekhidna.biocenter.helsinki.fi/dali_server

Most jobs finish within an hour, but if a queue builds up, then it takes longer.

DALI: Example result

Query: 1egfA

MOLECULE: EPIDERMAL GROWTH FACTOR;

Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, to pre-computed structural neighbours in the Dali Database, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.

Structural Alignment | Expand gaps | 3D Superimposition (Jmol Applet) | Reset Selection

Summary

No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
1:	1egf-A	99.9	0.0	53	53	100	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR;
2:	3egf-A	10.6	1.0	53	53	100	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR;
3:	3ca7-A	4.8	2.0	46	50	35	PDB	MOLECULE: PROTEIN SPITZ;
4:	1mox-D	4.5	3.0	47	48	32	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
5:	3c9a-C	4.4	2.0	44	48	36	PDB	MOLECULE: PROTEIN GIANT-LENS;
6:	3c9a-D	4.4	2.1	45	48	36	PDB	MOLECULE: PROTEIN GIANT-LENS;
7:	1ivo-C	4.3	2.7	44	47	61	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
8:	1mox-C	4.2	3.1	47	49	30	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
9:	1ivo-D	4.2	2.7	44	47	61	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
10:	1jl9-A	4.1	2.2	41	42	71	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR;
11:	1xdt-R	3.9	2.0	40	41	33	PDB	MOLECULE: DIPHTHERIA TOXIN;
12:	1bf9-A	3.7	2.4	39	41	33	PDB	MOLECULE: FACTOR VII;
13:	2vj3-A	3.7	2.9	41	120	32	PDB	MOLECULE: NEUROGENIC LOCUS NOTCH HOMOLOG PROTEIN 1;
14:	1epg-A	3.5	4.2	48	53	92	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR;
15:	1a3p-A	3.5	3.0	43	45	91	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR;
16:	1eph-A	3.4	4.5	48	53	92	PDB	MOLECULE: EPIDERMAL GROWTH FACTOR;
17:	1j9c-L	3.4	3.2	40	95	33	PDB	MOLECULE: TISSUE FACTOR;
18:	3ela-L	3.3	3.1	40	95	33	PDB	MOLECULE: COAGULATION FACTOR VII LIGHT CHAIN;
19:	1hae-A	3.3	3.1	48	63	27	PDB	MOLECULE: HEREGULIN-ALPHA;

DALI: Example result

Pairwise Structural Alignments

Notation: three-state secondary structure definitions by DSSP (reduced to H=helix, E=sheet, L=coil) are shown above the amino acid sequence. Structurally equivalent residues are in uppercase, structurally non-equivalent residues (e.g. in loops) are in lowercase. Amino acid identities are marked by vertical bars.

No 1: Query=1egfA Sbjct=1egfA Z-score=99.9

back to top

```
DSSP  LEELLLLLLLLLLLLLLEEEEELLLLLEEEELLLLLLLLLLLLLLLLLLL
Query NSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR   53
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct NSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR   53
DSSP  LEELLLLLLLLLLLLLLEEEEELLLLLEEEELLLLLLLLLLLLLLLLLLL
```

No 2: Query=1egfA Sbjct=3egfA Z-score=10.6

back to top

```
DSSP  LEELLLLLLLLLLLLLLEEEEELLLLLEEEELLLLLLLLLLLLLLLLLLL
Query NSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR   53
ident ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct NSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR   53
DSSP  LEELLLLLLLLLLLLLLEEEEELLLLLEEEELLLLLLLLLLLLLLLLLLL
```

No 3: Query=1egfA Sbjct=3ca7A Z-score=4.8

back to top

```
DSSP  -LEELLLLLLL-LLLLLLLLEEEEELL--LLLEEEELLLLLLLLLLLLLllllll
Query -NSYPGCPSSY-DGYCLNGGVCMHIES--LDSYTCNCVIGYSGDRCQTRDlrwwelr   53
ident        ||       ||||  |       |  | | | ||  | ||
Sbjct tFPTYKCPETFdAWYCLNDAHCFAVKIadLPVYSCECAIGFMGQRCEYKE-------   50
DSSP  lLLLLLLLHHHhHHLLLLLLEEEEEEEllEEEEEEELLLLEELLLLLEEL-------
```

Unterlagen zur Vorlesung

http://www.bpc.uni-frankfurt.de/guentert/wiki/index.php/Teaching