Benchmarking and Extension of Protein3DFit

Nasir Mahmood

Benchmarking and Extension of Protein3DFit

- Introduction
  - Why Align Structures?
  - Structure Alignment: Issues & problem
- Method
  - Algorithm
  - Pitfalls
  - Extension
- Results
- Outlook

Why Align Structures?

Fundamental step in:

- **Analysis & Visualization**
- **Comparison**
- **Design**

- Useful for:
  - **Structure classification**
    - To predict protein folds.
    - Gold standard for sequence alignment.
  - **Structure prediction:**
    - Identifying "structural core".
  - **Function prediction**
    - Function of protein is determined by its structure.
  - **Drug discovery**
    - To pinpoint active sites accurately.

Structure Alignment: Issues

Theoretical Issues

- NP-hard geometric problem
- Heuristics needed
- No unique solution

Methodological Issues

Choices:

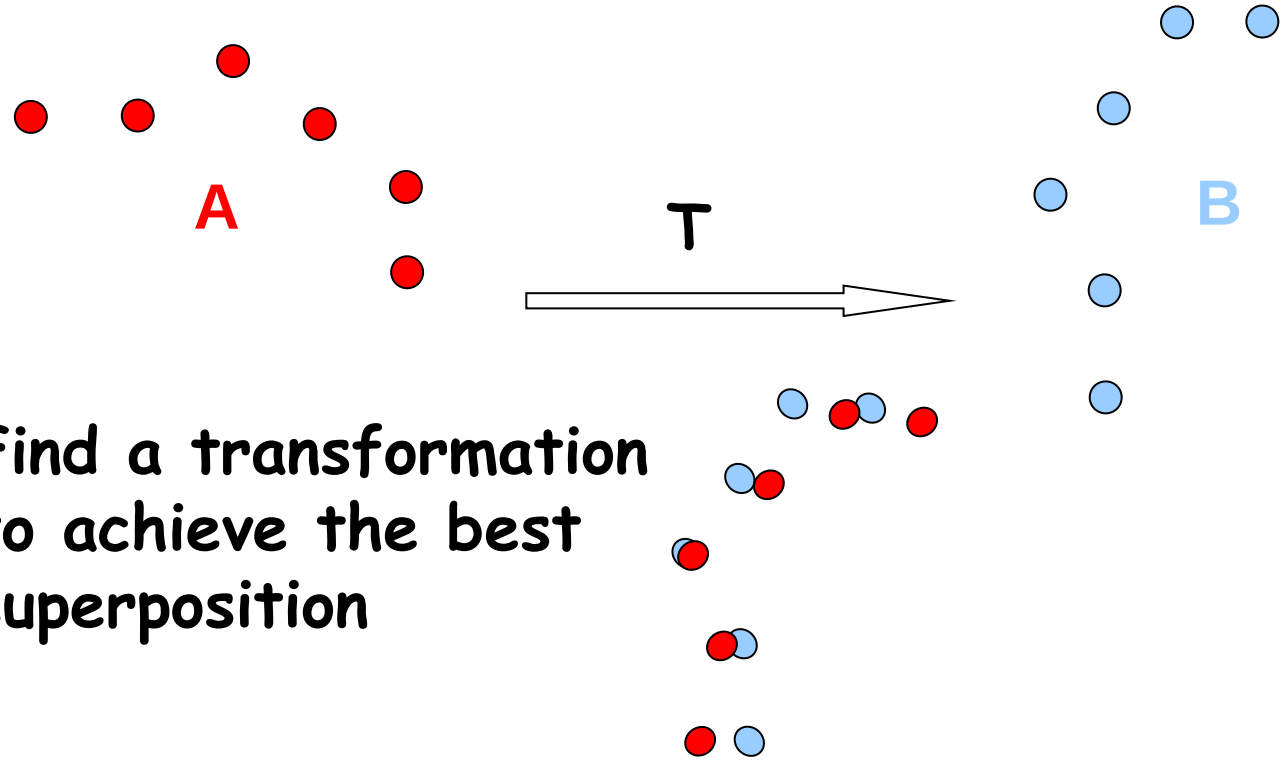
- Structure Representation
- Scoring function
- Search algorithm

Goals Desirable

- Automatic
- Discriminating
- Fast

Structure alignment Problem

Simple case – two closely related proteins with the same number of amino acids.



Find a transformation to achieve the best superposition

Structure Alignment Problem

Two sets of Points in 3D space:

$A = (a1, a2, \ldots, an)$

$B = (b1, b2, \ldots, bm)$

$A(P)$, $B(Q)$: Optimal subsets with $|A(P)|=|B(Q)|$,

Distance

$$\min_T \{ d(A(P) - T(B(Q))) \}$$

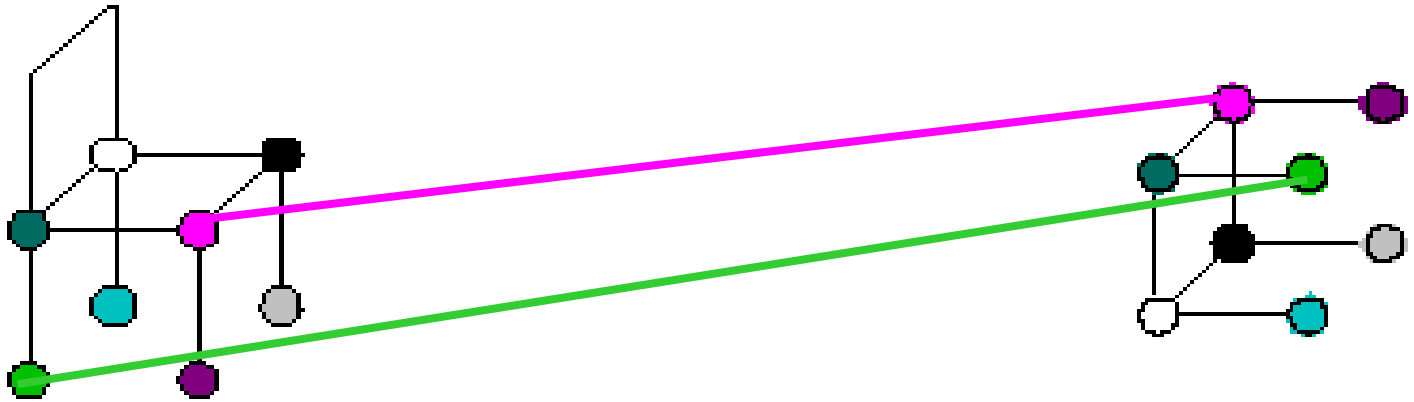
Optimal rigid body transformation
(calculated by Diamond (1998) method)

Two Subproblems

1. Find correspondence set
2. Find optimal transformation

   (protein superposition problem)

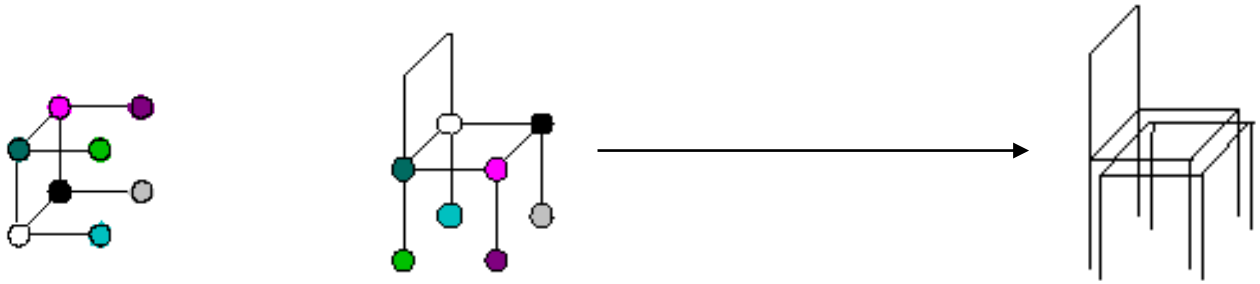
Correspondence set



- Correspondence is a key problem.
- Exhaustive search intractable.
- Heuristics are applied.

Superposition

- Two sets of points
  $A = \{a1,\ldots,an\}$ and $B = \{b1,\ldots,bn\}$



- Correspondence pairs (ai, bi)
- Find T

$$\min_T \{ RMSD(A, T(B)) \}$$

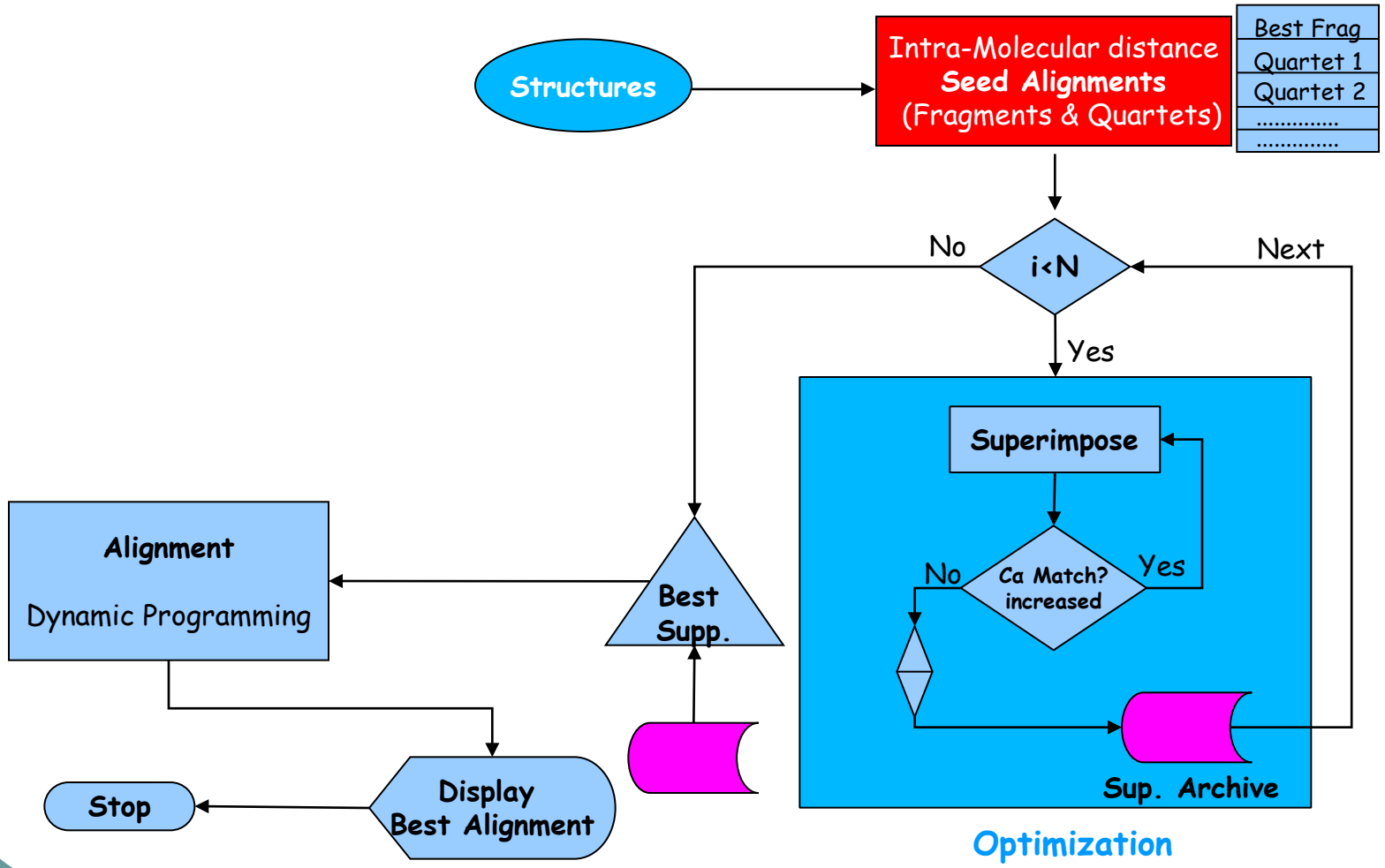
Aim of the project

- To make algorithm more efficient
  - debugging
  - optimization
  - extension
  - imporve quality of alignments
  - organization of code

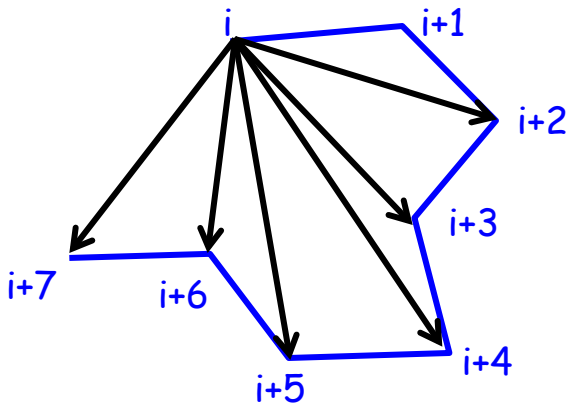
Benchmarking and Extension of Protein3DFit

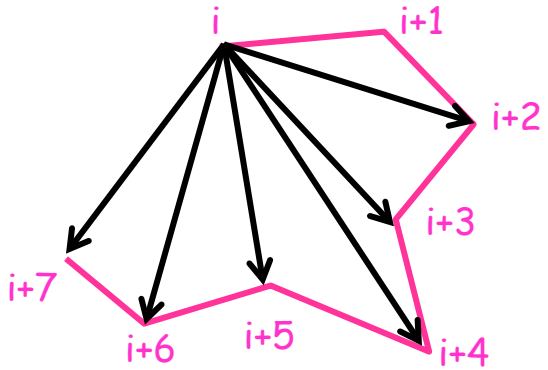
- Introduction
  - Why Align Structures?
  - Structure Alignment: Issues & problem
- Method
  - Algorithm
  - Pitfalls
  - Extension
- Results
- Outlook

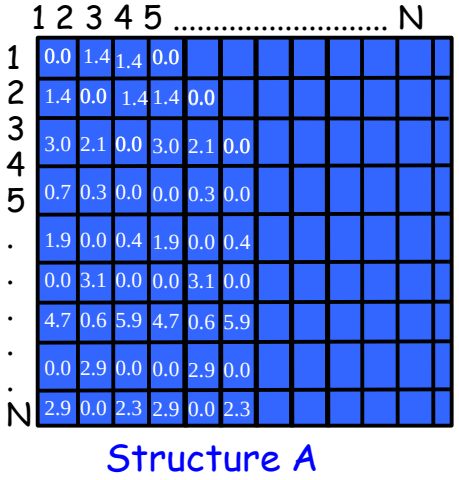
Algorithm


Structures
Intra-Molecular distance
Seed Alignments
(Fragments & Quartets)
Best Frag
Quartet 1
Quartet 2
.....
.....
No
i<N
Next
Yes
Superimpose
No
Ca Match?
increased
Yes
Sup. Archive
Best
Supp.
Alignment
Dynamic Programming
Display
Best Alignment
Stop
Optimization

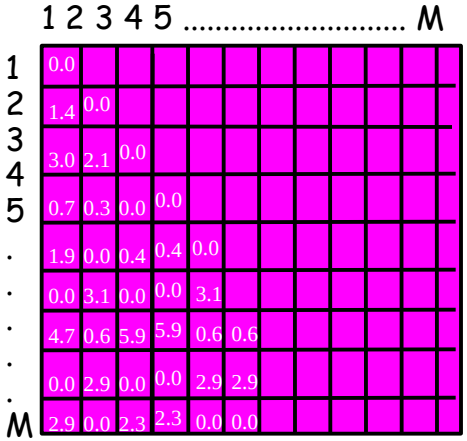
Intra-molecular Distance Matrices





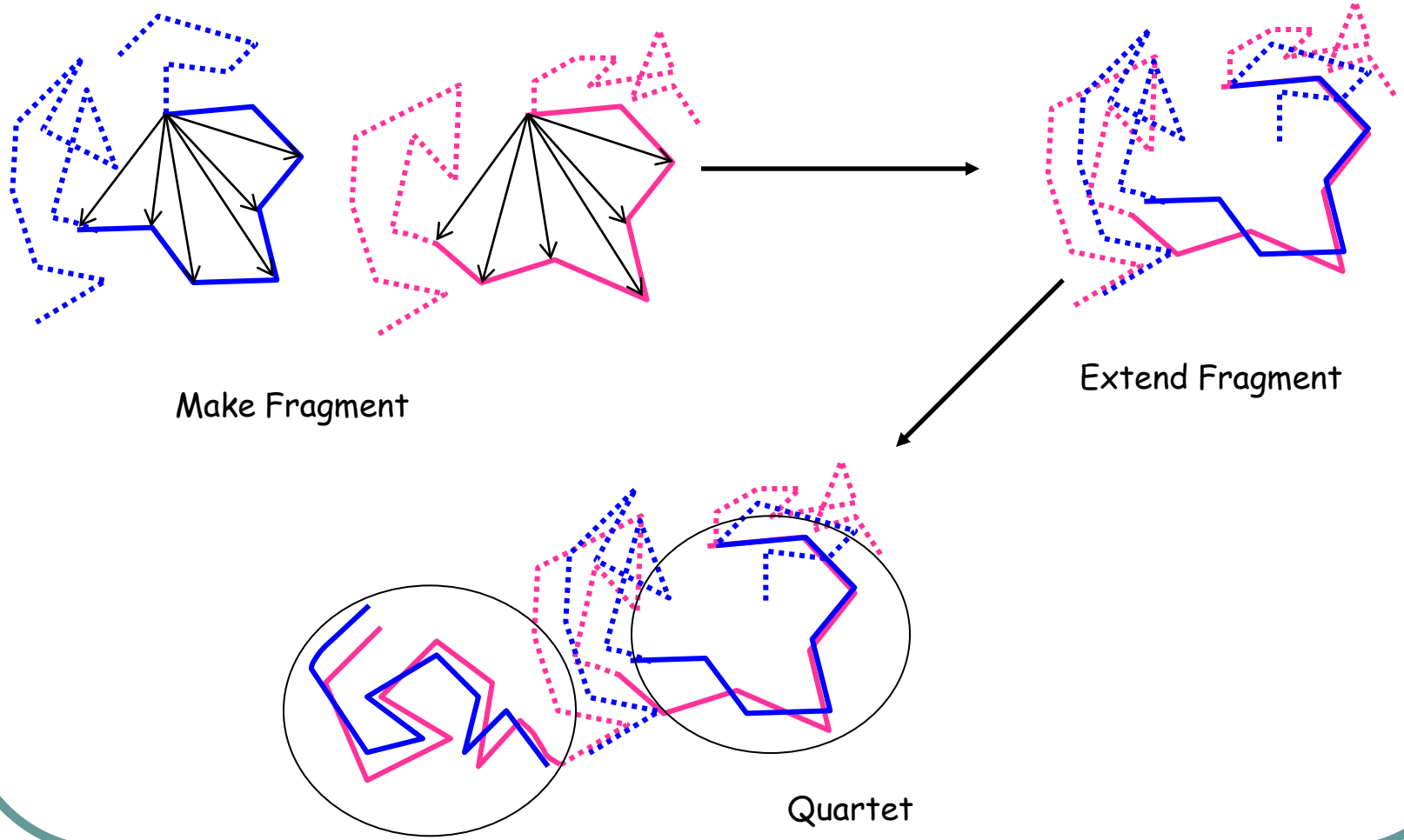


Structure A

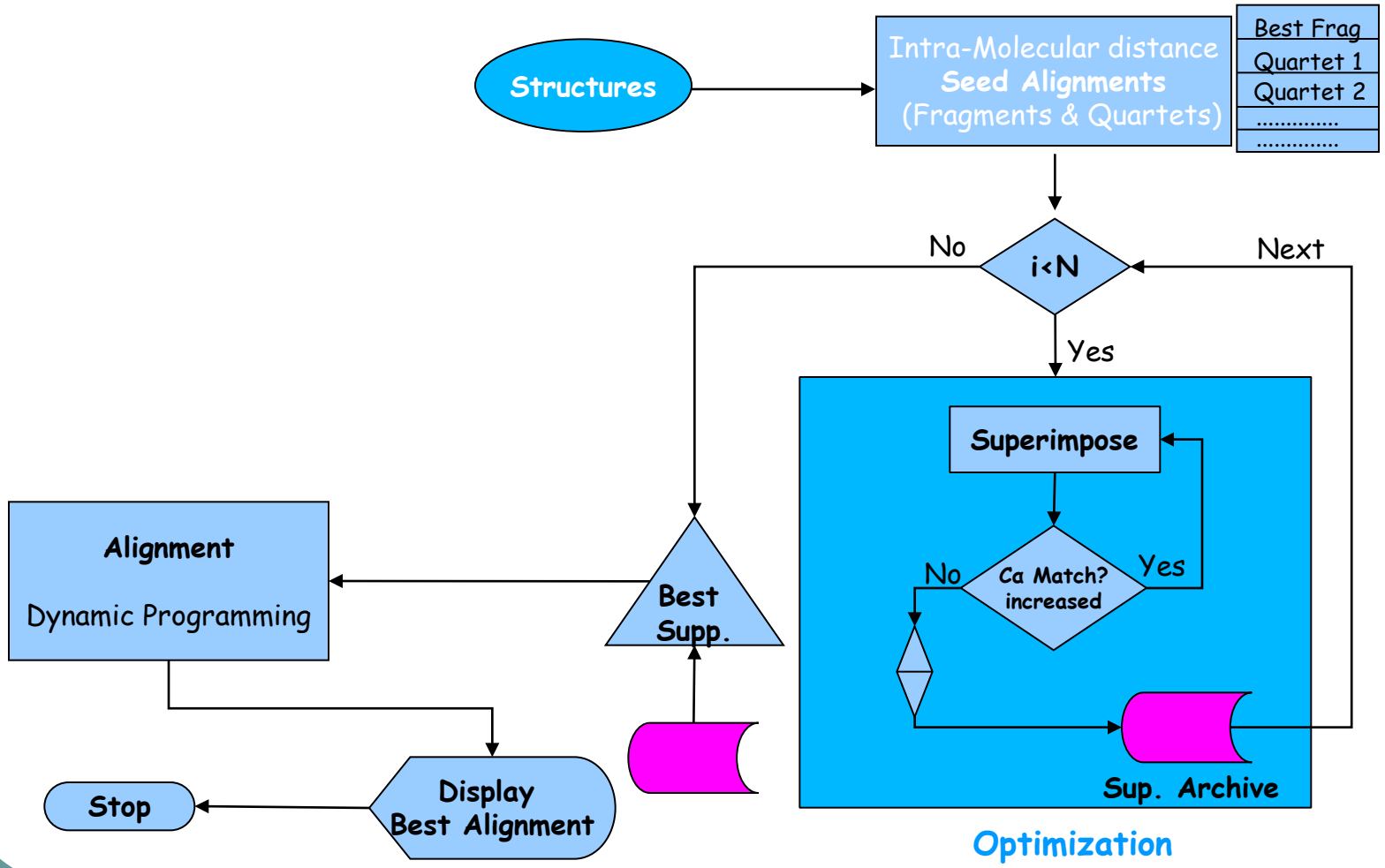


Structure B

Seed Alignments

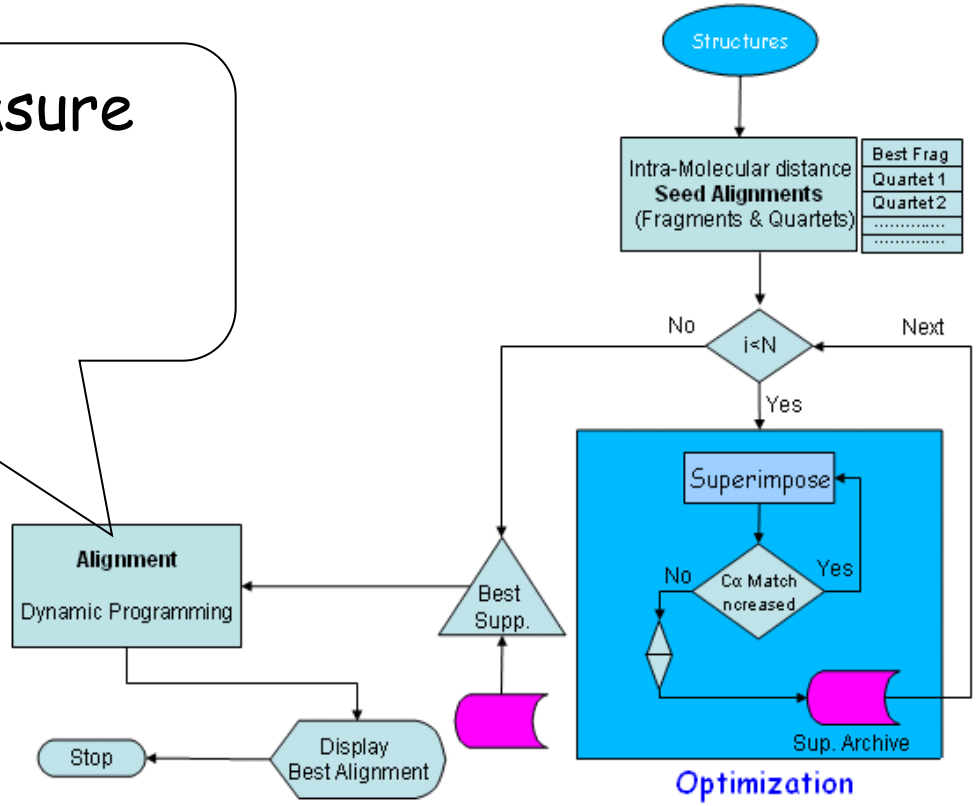


Algorithm

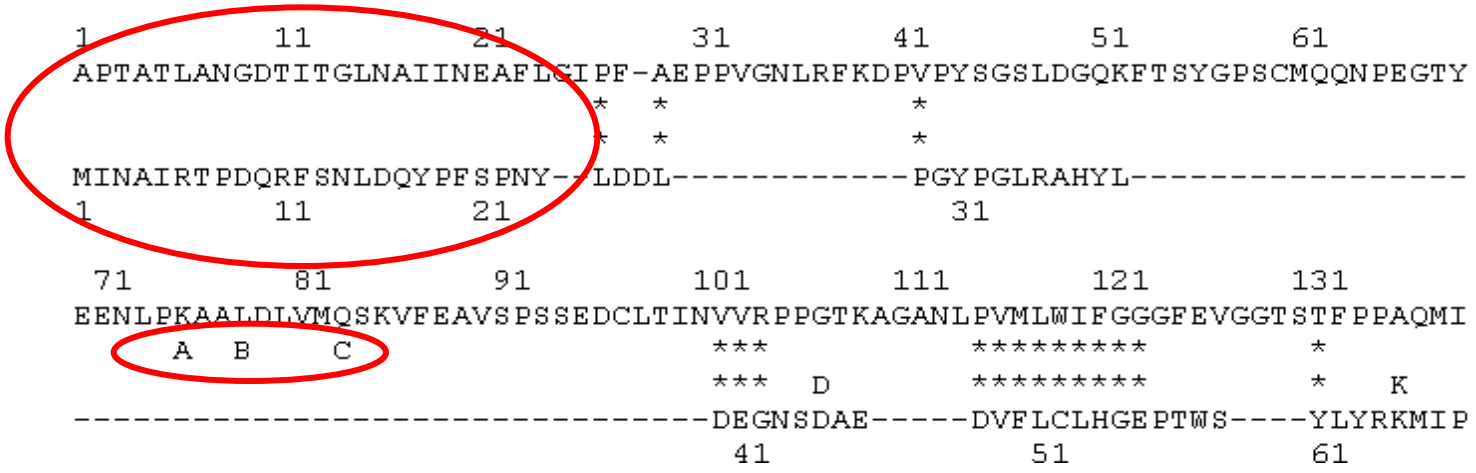

Structures
Intra-Molecular distance
Seed Alignments
(Fragments & Quartets)
Best Frag
Quartet 1
Quartet 2
.....
.....
No
i<N
Next
Yes
Superimpose
No
Ca Match? increased
Yes
Sup. Archive
Optimization
Best Supp.
Alignment
Dynamic Programming
Display Best Alignment
Stop

Algorithm- Pitfalls

- Simple Similarity measure
  - 1: Match
  - 0: No match



Algorithm- Pitfalls



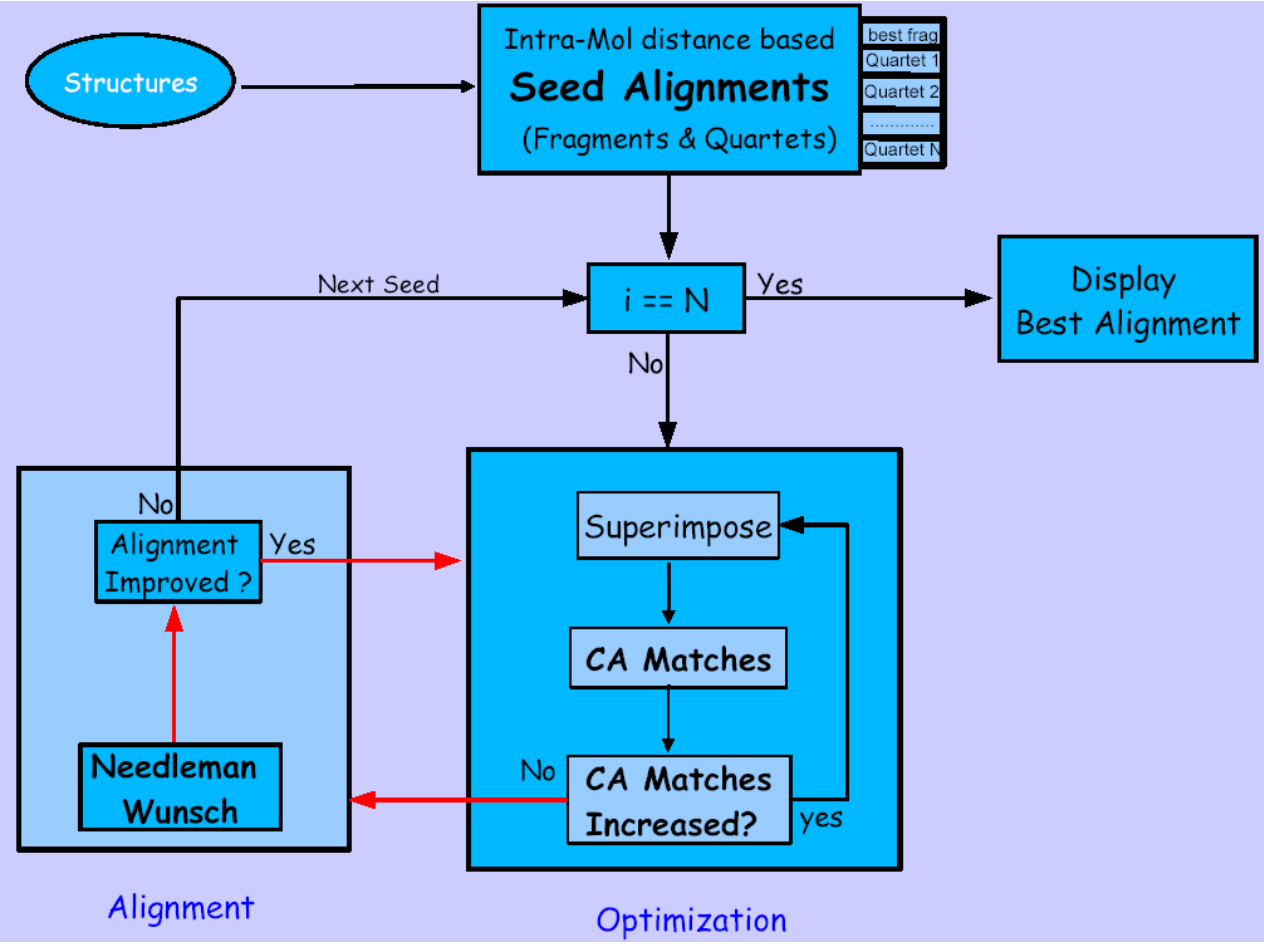
- Problems in aligning two structures low similarity.
- Alignments sometimes fragmented.
- Single matching residues not in sequential order (letters in the alignment).
- No optimization after dynamic programming.

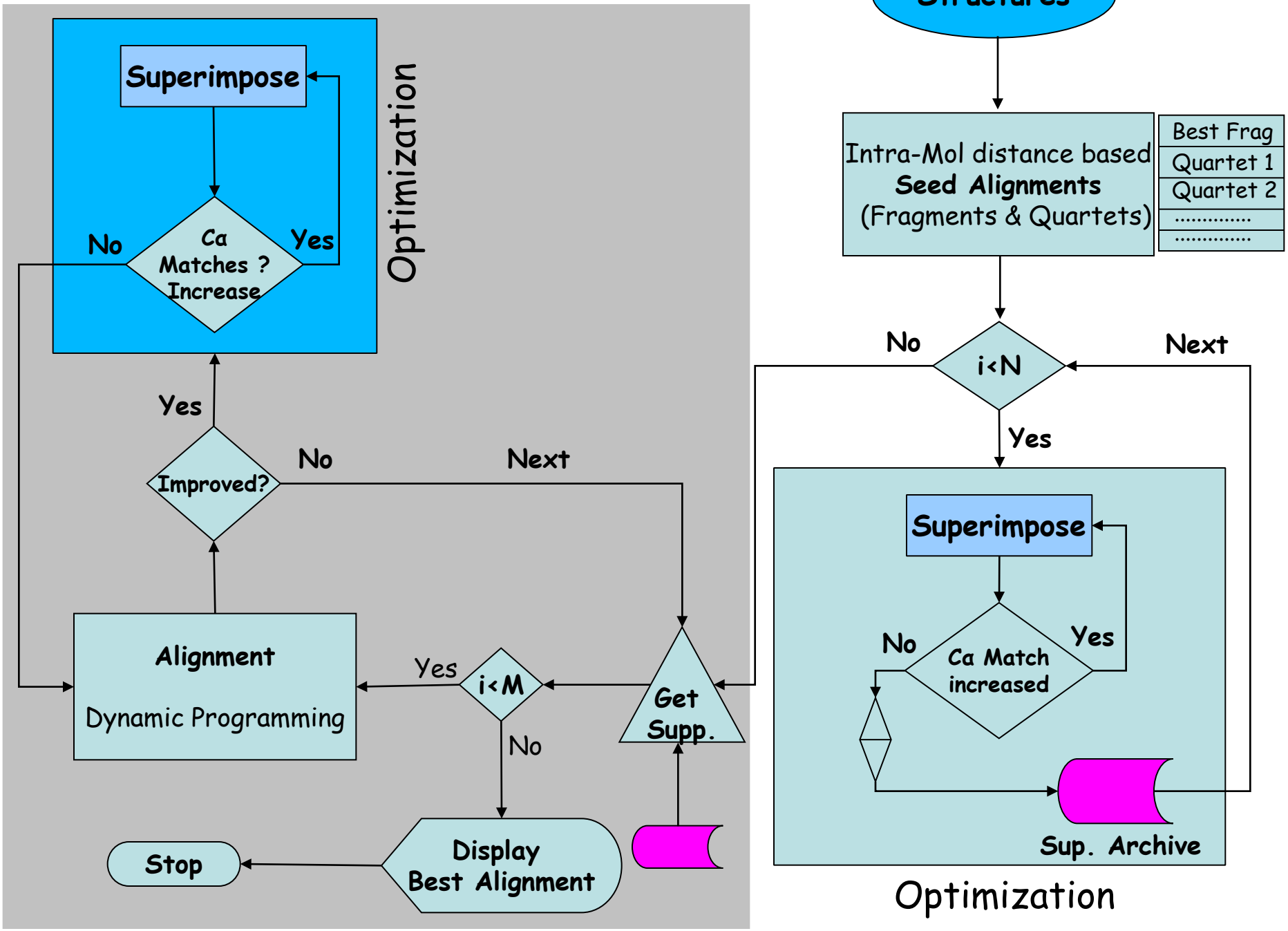
Algorithm- Pitfalls

No	Chain 1	Chain 2	Protein3DFit $N^A/r.m.s.d.(A^o)$	VAST $N^A/r.m.s.d.(A^o)$	DALI $N^A/r.m.s.d.(A^o)$	CE $N^A/r.m.s.d.(A^o)$
1	1FXI:A	1UBQ:_	**39/1.0**	48/2.1	_	_
2	1TEN:_	3HHR:B	**71/1.0**	78/1.6	86/1.9	87/1.9
3	3HLA:B	2RHE:_	**42/1.1**	_	63/2.5	85/3.5
4	2AZA:A	1PAZ:_	**52/1.2**	74/2.2	_	85/2.9
5	1CEW:I	1MOL:A	**54/1.2**	71/1.9	81/2.3	69/1.9
6	1CID:_	2RHE:_	**59/1.1**	85/2.2	95/3.3	94/2.7
7	1CRL:_	1EDE:_	**100/1.2**		211/3.4	187/3.2
8	2SIM:_	1NSB:A	**129/1.2**	284/3.8	286/3.8	264/3.0
9	1BGE:B	2GMF:A	**52/1.1**	74/2.5	98/3.5	94/4.1
10	1TIE:_	4FGF:_	**70/1.1**	82/1.7	108/2.0	116/2.9

Reference: 10 'difficult' structures from (Fischer et al.) obtained by Dali (Holm & Sander, 1993) , VAST (Madej et al., 1995) & CE (Shindyalov & Bourne, 1998)

Algorithm - Extension




Structures
Intra-Mol distance based
Seed Alignments
(Fragments & Quartets)
Best Frag
Quartet 1
Quartet 2
.....
.....
No
i<N
Next
Yes
Superimpose
No
Ca Match increased
Yes
Sup. Archive
Optimization
Superimpose
No
Ca Matches ? Increase
Yes
Optimization
Yes
Improved?
No
Next
Alignment
Dynamic Programming
Yes
i<M
No
Get Supp.
Stop
Display Best Alignment

Similarity Score

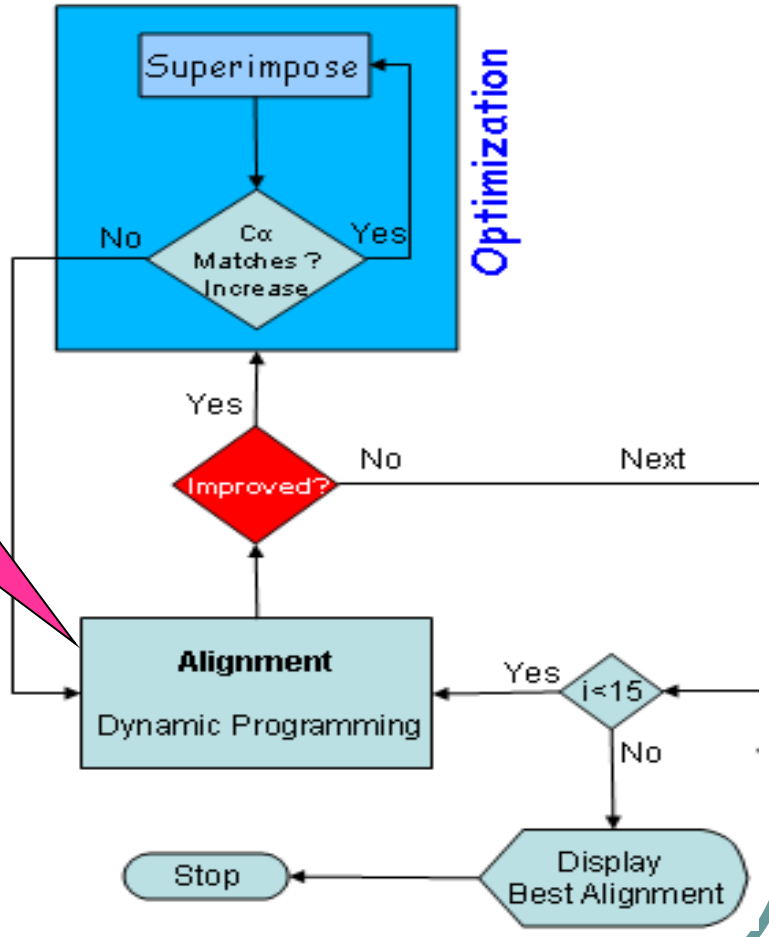
$$S_{ij} = \frac{M}{1 + (\frac{d_{ij}}{d_0})^2}$$

S_{ij} :similarity score used in DP

M :maximum score for a match

d_{ij} :distance

d_o :distance resulting in half-maximal similarity



Source: Gerstein Levitt, 1998

Quality Function

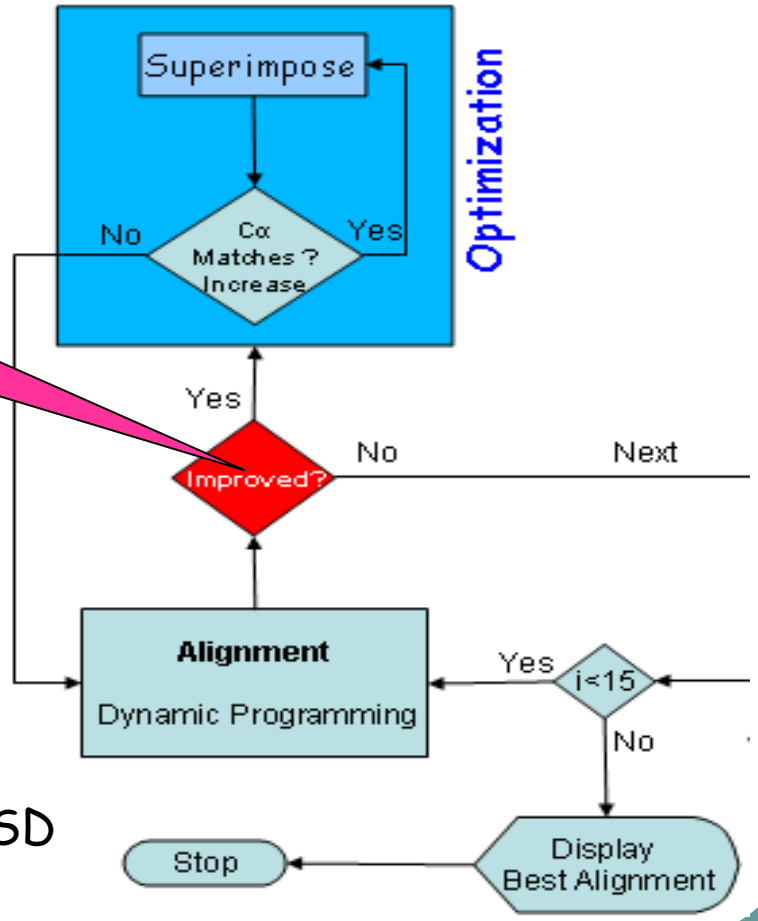
Q score, ratio of no. of aligned residues & RMSD

$$Q = \frac{N_{align}^2}{((1+(RMSD/R_0)^2)N_1N_2)}$$

N_{align} :number of matches (equivalent residues)

N_1, N_2 :length of structure 1 & 2 respectively

R_O :empirical parameter measuring Relative significance of N_{align} & RMSD



Source: Krissinel & Henrich (2004)

Benchmarking and Extension of Protein3DFit

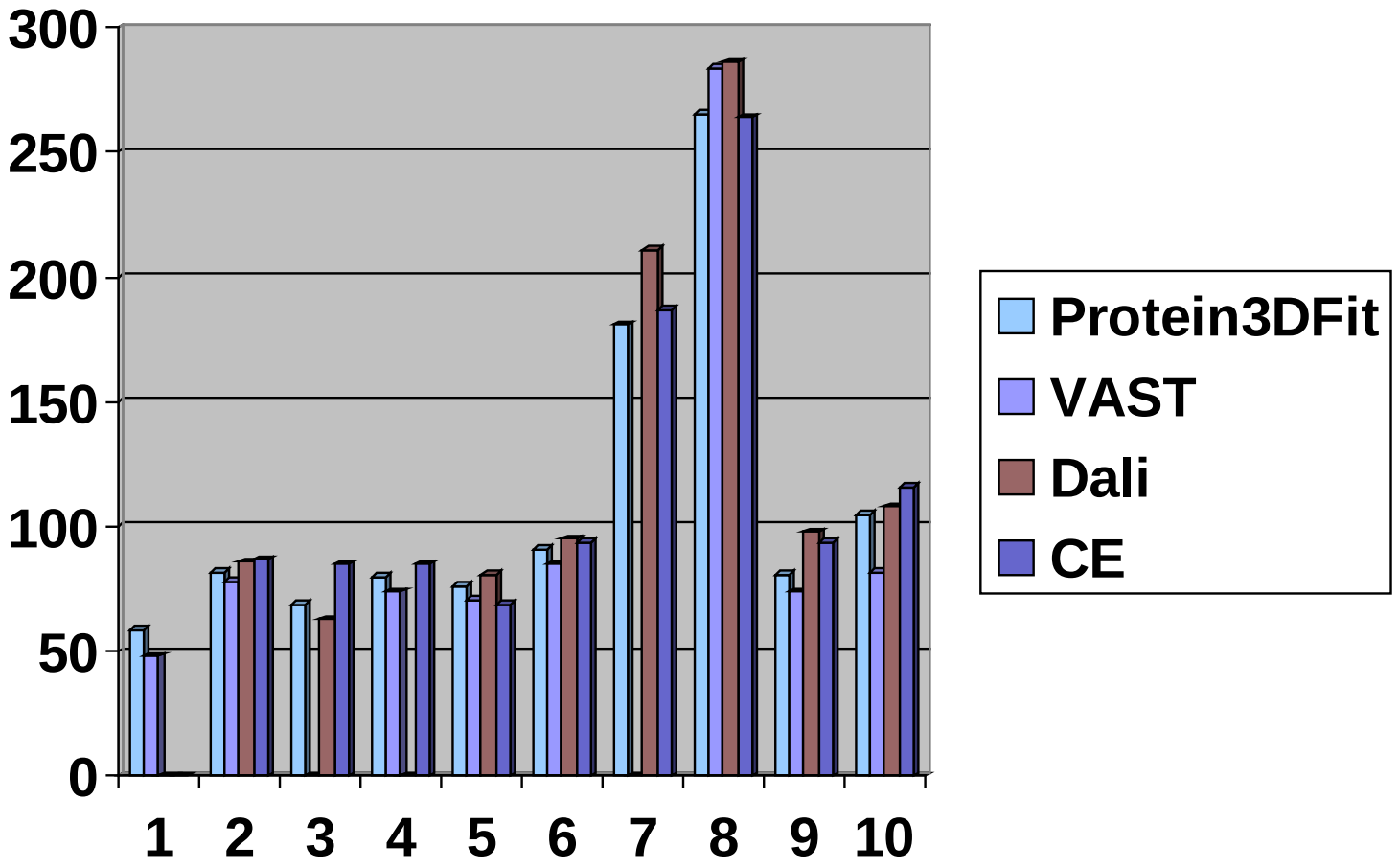
- Introduction
  - Why Align Structures?
  - Structure Alignment: Issues & problem
- Method
  - Algorithm
  - Pitfalls
  - Extension
- Results
- Outlook

Results

No	Chain 1	Chain 2	Protein3DFit $N^A/r.m.s.d.(A^o)$		VAST $N^A/r.m.s.d.(A^o)$	DALI $N^A/r.m.s.d.(A^o)$	CE $N^A/r.m.s.d.(A^o)$
1	1FXI:A	1UBQ:_	59/2.9	39/1.0	48/2.1	_	_
2	1TEN:_	3HHR:B	82/1.7	71/1.0	78/1.6	86/1.9	87/1.9
3	3HLA:B	2RHE:_	69/3.4	42/1.1	_	63/2.5	85/3.5
4	2AZA:A	1PAZ:_	80/2.3	52/1.2	74/2.2	_	85/2.9
5	1CEW:I	1MOL:A	76/2.1	54/1.2	71/1.9	81/2.3	69/1.9
6	1CID:_	2RHE:_	91/2.6	59/1.1	85/2.2	95/3.3	94/2.7
7	1CRL:_	1EDE:_	181/3.2	100/1.2		211/3.4	187/3.2
8	2SIM:_	1NSB:A	265/3.2	129/1.2	284/3.8	286/3.8	264/3.0
9	1BGE:B	2GMF:A	81/3.0	52/1.1	74/2.5	98/3.5	94/4.1
10	1TIE:_	4FGF:_	105/2.6	70/1.1	82/1.7	108/2.0	116/2.9

source: 10 'difficult' structures from (Fischer et al.) obtained by Dali (Holm & Sander, 1993) , VAST (Madej et al., 1995) & CE (Shindyalov & Bourne, 1998).

Results


300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
Protein3DFit
VAST
Dali
CE

3Dfit-ext

```
CA Matches: 59 (61.46%/77.63%), RMS-Deviation:2.95, Q-Value:0.84  Gaps: 44
1FXI:A   96 residues  ->  1UBQ:_   76 residues

          11          21          31          41          51          61
 YKVTLKT-PDGDNVITV-PD-DE--YILDVAEEEGL-DLPYSCRAGACSTCAGKLVSGPAPDEDQSFL
 ******* ********* **  *  ********* * ****        ******
 ******* ********* **  *  ********* * ****        ******
 MQIFVKTLTGKTITLEVEPSD-TIENVKAKIQDK-EGIPPD--------QQRLIF-------------
 1         11        21         31                  41

     71         81             91
DDDQIQAGYIL-TCVAYPT-----GDCVIETHK-E
    *****    ******      ********  *
    *****    ******      ********  *
-----AGKQL-E-DGRTLSDYNIQKESTLHLV-LR
           51         61         71
```

CE

```
1FXI:A     5/6     VTLKTPD-GDNVITVPDDEY---ILDVAEEEGLDLPYSCRAGACSTCAGKLVSGPAPDED
1UBQ:_     3/4     IFVKTLTGKTITLEVEPSDTIENVKAKIQD-KEGIPPD-------QQRLIFA--------

1FXI:A    61/62    QSFLDDDQIQAGYILTCVAYPT-----GDCVIETHKEEAL
1UBQ:_    47/48    -----------GKQLEDGRTLSDYNIQKESTLHLVLRLRG
```

Rmsd = 3.8Å Z-Score = 3.1
Sequence identity = 9.4%
Aligned/gap positions = 64/36

3Dfit-old

```
pdb1fxi   96 residues  ->  pdb1ubq   76 residues
matching Ca:  39  ( 40.62% / 51.32% )
rms deviation: 1.027523  min. length:  6

1          11          21          31          41          51          61
ASYKVTLKTPD-GDNVITVPDDE---YILDVAEEEGL-DLPYSCRAGACSTCAGKLVSGPAPD-EDQSFL
   * ****   ******        ********* * * A*          ***   B cc *
   * ****   ******  F   E ********* * *  *A          ***    D   *cc
M--QIFVKTLTGKTITLEVEPSDTIENVKAKIQDK-EGIPPDQQR--------LIFAGKQLEDGRTLSDY
1           11          21          31          41                  51

      71         81         91
DDDQIQAGYILTCVAYPTGDCVIETHKEEALY
         D    E  F   *****
 B                  *****
NIQKES---------------TLHLVLRLRGG
 61                       71
```

Dali

No 1: Query=mol1A Sbjct=mol2 Z-score=4.0

```
DSSP  llEEEEEEEL.LEEEEEEELLLL...LHHHHHHHHLLLLLLllllllllLLEEEEeellll
Query asYKVTLKTP.DGDNVITVPDDE...YILDVAEEEGLDLPYscragacsTCAGKLvsgpa   56
ident        ||         |                    |
Sbjct ..MQIFVKTLtGKTITLEVEPSDtieNVKAKIQDKEGIPPD........QQRLIF.....   45
DSSP  ..LEEEEEELlLLEEEEELLLLLlhhHHHHHHHHHHHLLLHH........HEEEEE.....


DSSP  lllllllllllllLLEEEhHHLLL.....LLLEEEELllllll
Query pdedqsfldddqiqAGYILtCVAYP.....TGDCVIETHkeealy   96
ident
Sbjct .............aGKQLE.DGRTLsdyniQKESTLHLVlrlrgg   76
DSSP  .............lLEELL.LLLLLhhhllLLLLEEEEEellllll
```

3Dfit-ext

```
CA Matches:  69  (69.70%/60.53%), RMS-Deviation:3.47, Q-Value:0.82  Gaps: 55
3HLA:B   99 residues  ->  2RHE:_  114 residues

         11          21          31        41
     KIQVYSRHPAE-NG-KSNFLN-C-YVSGF--HPSDIEVDLLKN-----GERI-E-----K-----
     **** *  ***  * *****     ****  *** ********     ****       *
     **** *  ***  * *****     ****  *** ********     ****       *
     VLTQ-PPSASGT-PGQRVTI-S-C-TGSATDIGS-NSVIWYQQVPGKAPKLLI-YYNDLLPSGVS
              11          21          31        41         51

     51         61         71             81        91
---VEHS-DLSFSKDWSFY-LLYYTEFT-----PTEKDEYACRVNHVTLSQPKIV-KWD-R
   ***   *        *  *******     **  ****************** **  *
   ***   *        *  *******     **  ****************** **  *
DRFSAS-K-S--------GT-SASLAISGLESEDE--ADYYCAAWNDSLDEPGFGGGT-KL
61                  71        81          91        101
```

CE

```
3HLA:B     4/5     TPKIQVYSRHPAENGKSNFLNCYVSGFHP-SDIEVDLLKN-----GERIE----------
2RHE:_     3/4     VLTQPPSA---SGTPGQRVTISCTGSATDIGSNSVIWYQQVPGKAPKLLIYYNDLLPSGV

3HLA:B    48/49    ---KVEHSDLSFSKDWSFYLLYYTEFTPTEKDEYACRVNHV--TLSQPKIVKWDR
2RHE:_    60/61    SDRFSASKSG-------TSASLAISGLESEDEADYYCAAWNDSLDEPGFGGGTKL
```

Rmsd = 3.4Å Z-Score = 3.7
Sequence identity = 2.4%
Aligned/gap positions = 84/31

3Dfit-old

```
1_pdb3hla   99 residues  ->  pdb2rhe  114 residues
matching Ca:  42  ( 42.42% / 36.84% )
rms deviation: 1.166318  min. length:  7

1                                    11
IQR--------------------------TPKIQVYSRHPAE--------------------------
                             ** ******
F     E           D C B A    ** ******                          ooo N
ESVLTQPPSASGTPGQRVTISCTGSATDIGSNSVIWYQQVPGKAPKLLIYYNDLLPSGVSDRFSASKSGT
1         11        21        31        41        51        61

               21        31        41        51        61         71
-----------NGKSNFLNCYVSGFHPSDIEVDLLKNGERIEKVEHSDLSFSKDWSFYLLY-YTEFTPTE
           *   ********      A B C D       E  F   *    G   ** ****hhI
 M L K J I *hh ******** G                          *        ** ****
SASLAISGLESEDE-ADYYCAAWNDS-----------------------LDEPG----FGGGTKLTVLG
71        81         91                                    101       1

      81        91
KDEYACRVNHVTLSQPKIVKWDRDM
J K L M N       ooo

QPK
11
```

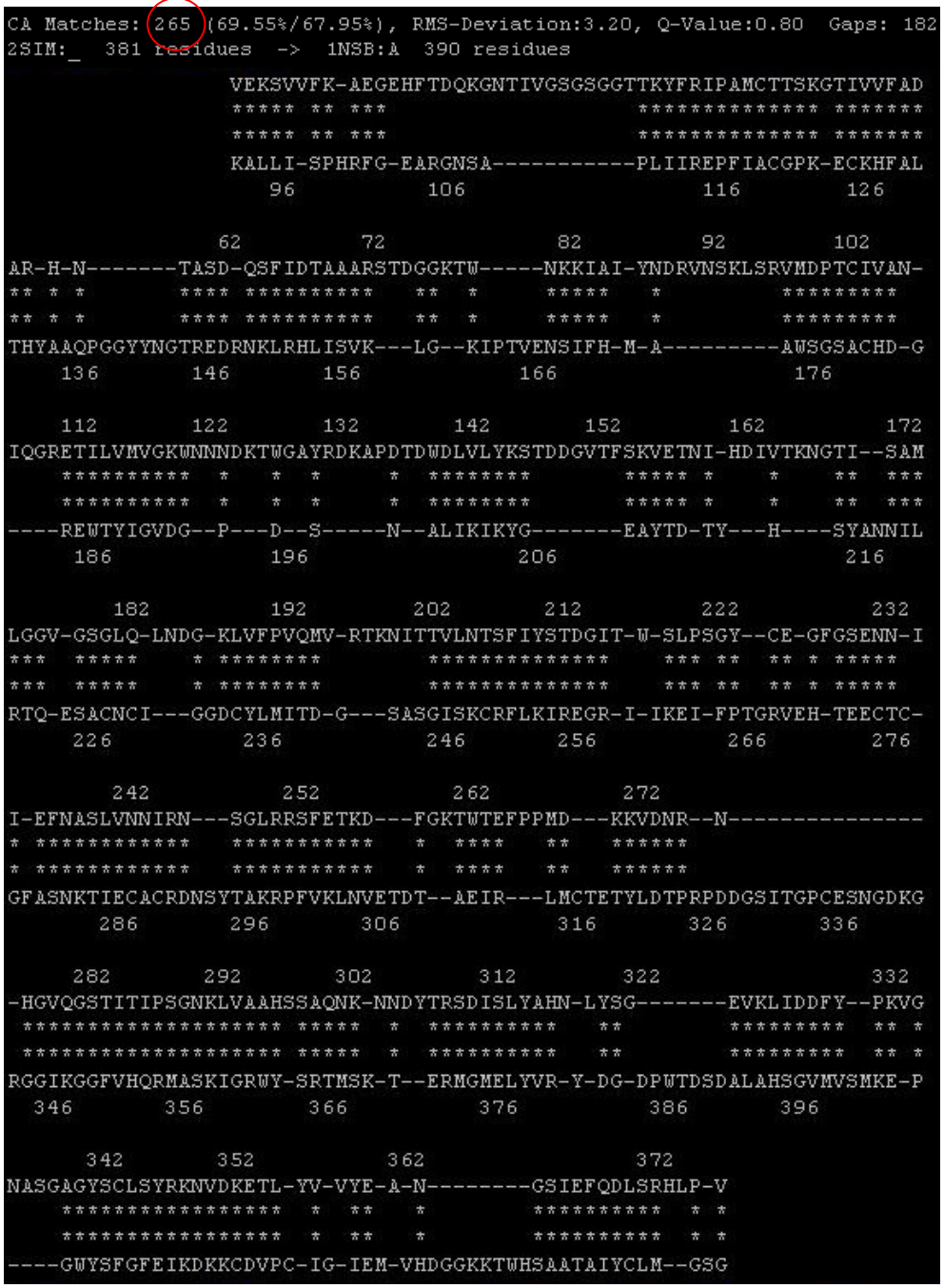
Dali

No 1: Query=mol1B Sbjct=mol2 Z-score=5.4

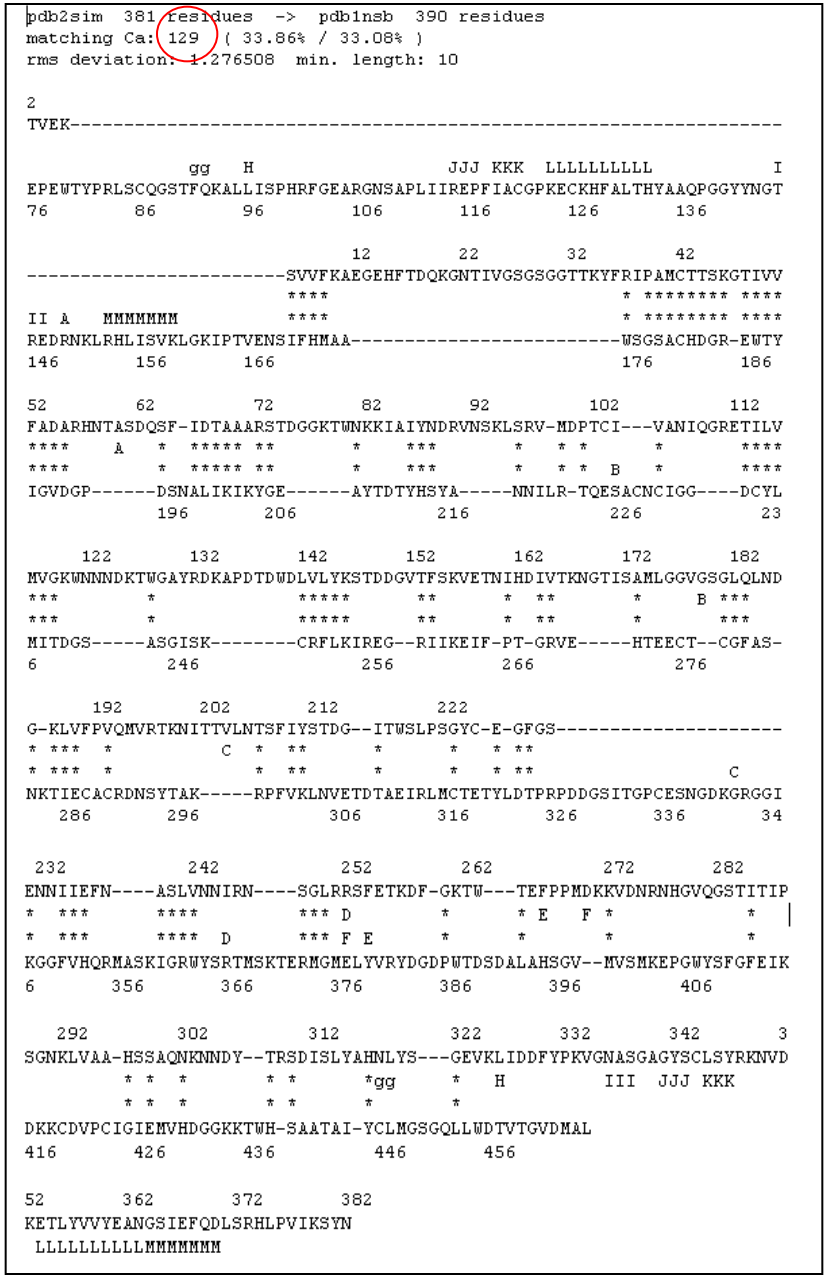
```
DSSP  llllleEEEEEellLLLLLLLLeEEEEEEEEEllLEEEEEEEL.....LEELL........
Query iqrtpKIQVYsrHPAENGKSNFLNCYVSGFhpsDIEVDLLKN.....GERIE........   47
ident                    |      |  |        |
Sbjct ...esVLTQPpsASGTPGQRVTISCTGSATdigSNSVIWYQQvpgkaPKLLIyyndllps   57
DSSP  ...llLLLLLleEEELLLLLEEEEEELLLLllLLLEEEEELllllLEEEElllllll


DSSP  ......LLEELlleellllleEEEEEEEE.llLLLL..LEEEEEEELLLLLLEEE.ELLL
Query ......KVEHSdlsfskdwsfYLLYYTEF.tpTEKD..EYACRVNHVTLSQPKIV.KWDR   97
ident           |                              |   | |      | |
Sbjct gvsdrfSASKS..........GTSASLAIsglESEDeaDYYCAAWNDSLDEPGFGgGTKL  107
DSSP  llllleEEEEE..........LLEEEEEEllLHHHleEEEEEEEELLLLEEEELlLEEE


DSSP  .....ll
Query .....dm   99
ident
Sbjct tvlgqpk  114
DSSP  eellllll
```



3Dfit-ext



3Dfit-old

Outlook / Possible Extensions

- Use of secondary structure information in alignment step.
- Update Protein3DFit webserver
- Multiple structure superposition
- Alignment of a protein structure against PDB
- ……

Acknowlegdments

- Prof. Dietmar Schomburg
- Pascal Benkert
- Structural Biology Group & CUBIC

Questions?

Thanks a lot for your patience!