

Bioinformatics Course September 2002

Molecular Modelling

Nathan Hall

Ludwig Institute for Cancer Research



What is Molecular Modelling and Why?

Answer Biologically Relevant Questions

- * Secondary Structure prediction
- * Comparative or Homology modelling
- * Threading, tertiary structure prediction
- * Protein Simulation
- * Docking
 - Protein-Protein
 - Protein-Ligand
- * Drug Design



Secondary Structure Prediction

Predict α -helix and β -sheet regions in a protein sequence

H=Helix,

E=Extended (β -sheet)

C=Coil (no secondary structure - loop regions or random coil)

- Improved results by aligning a family of sequences, then making predictions based on the sequence profile.
 - similar sequences will have conserved secondary structure

eg **PSIpred** Jones, D. T. (1999) *J. Mol. Biol.* 292, 195-202

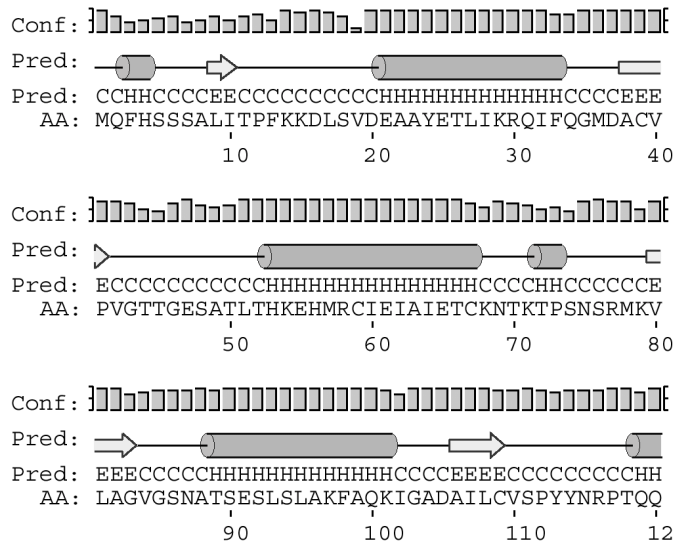
<http://insulin.brunel.ac.uk/psipred/> (e-mail server)

- 70-80 % accuracy

- initially generates a psiBlast alignment & profile



Sample PSIpred output



Homology/Comparative Modelling

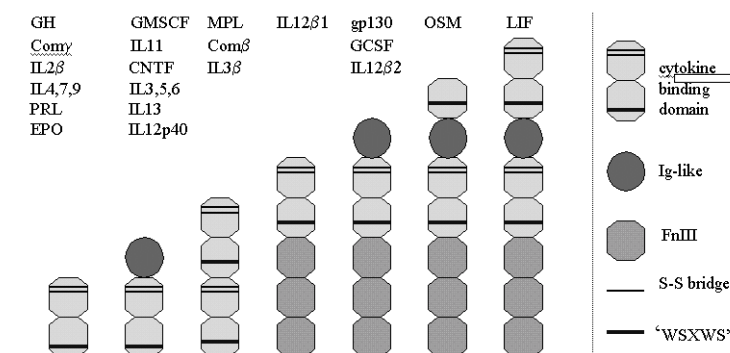
- Protein with *similar* sequences tend to adopt the same general fold
- Able to predict the structure of a given sequence based on structural template with a similar sequence
- Provide functional predictions and explanations
- Typically requires > 25-30 % sequence identity

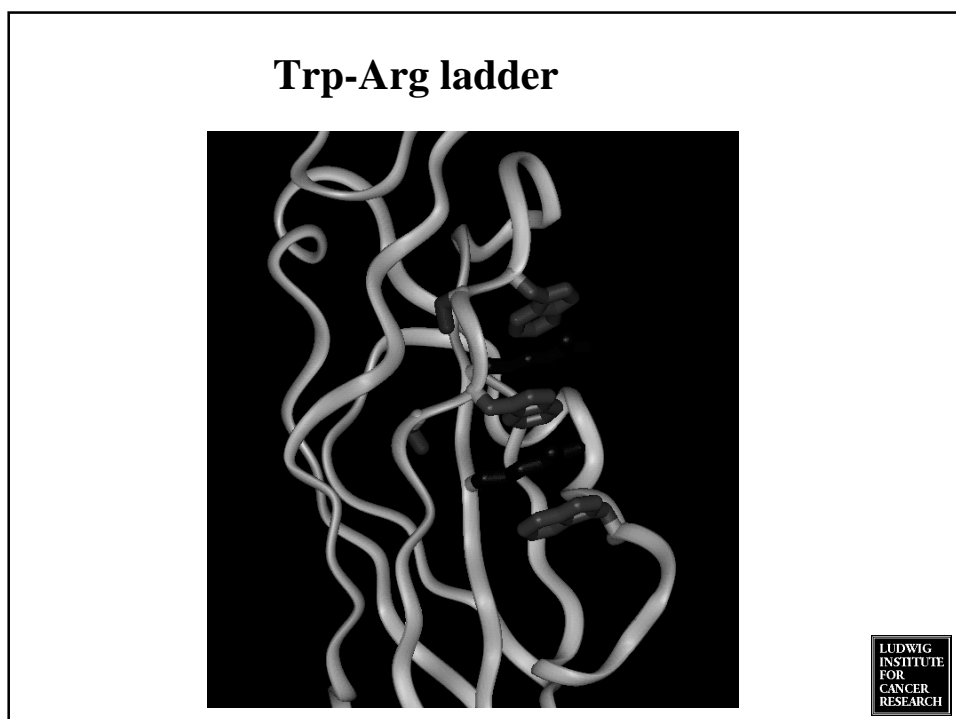
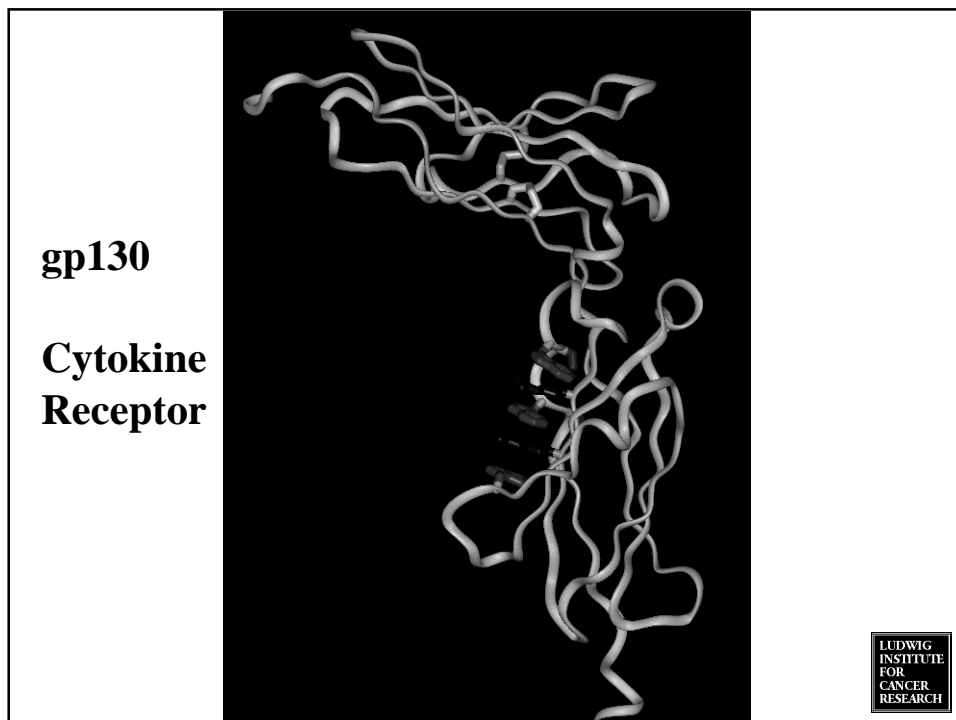


eg cytokine receptors



Class I Cytokine Receptors





Structural Sequence Alignment of Cytokine Receptors

```

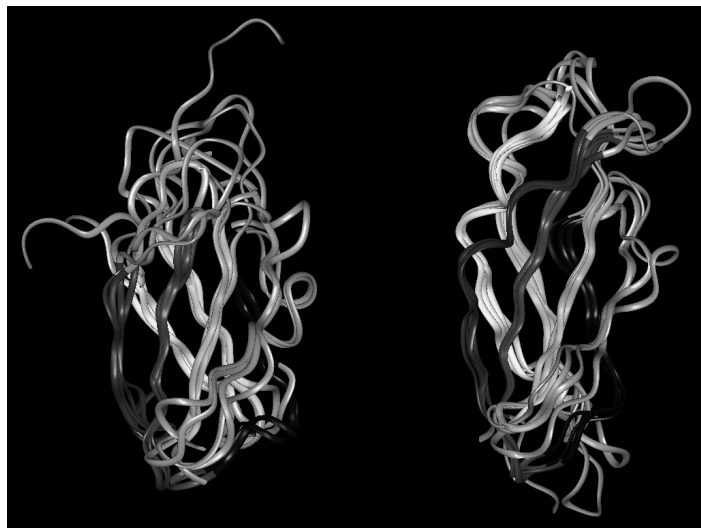
gp130  P EK KNL-SCIIVNE-GKKMRCE DGGRETHL--ETNFTLKSEWAT-----HKPADCKAKADTP-----TSCQVDY
mGCSFR P AS SNL-SCLMHLTNSLVCQ EPGPETHL--PTSFILKFSRSDCQ---YQGDITPDCVAKKR-----QNNCSLFR
rPLR   GK -EIHKCRSPD-KETFTCW NPCTDGL--PTNYSILYSKEG-----EKTTY-ECPDYKTSG-----PNSCFPSK
GHR    E--KFTKCRSPE-RETFTSCH TDEVHHgknlGPIQLFYTRRNTqewtgEWTQEWK-ECPDYVSAG-----ENSCYFNS
EPOR   AARG EEL-ICETER-LEDLVCF EEAASAGVg-PGNYSFSYQLE-----DEPWK-LCRLHQAPAR----GAVRFWCSLPT
IL4R   FKVLQEP-FCVSDY-MSISTCF KMNQPTNC--STELRLLYLQVFL-----SEAH-TCIPENNG-----GAGCVCHLLM
IL12R  KTFL-RCEAKNYSGRFTCW LTTTIS----TDLTFSVKSSRGssd-----pqGVTCGAATLSAERVRGDNKEYEYSVECCQE
          1A          1B          1C          1C'          1E

gp130  -----STVYFVNIEVWVEAENALGKVTSDHINFDPVYKVK N HNLVINSEEL-----SSILKLT TNPS----IKSV
mGCSFR -----KNLLLYQYMAIWVQAEENMLGSSSEPALCLDFMDVVKLE M-LQALDIgpdvvhqPGCLWLS KPWK----PSEY
rPLR   -----QYTSIWKIYIITVNATNQMGSSSDPLYVDVYIIVE E RNLTLVVKqk--dKRTYLWVK SPPTITDVKTGW
GHR    -----SFTSIWPIYCIKLTSN----GGTVDEKCFSVDEIVQ D TALNWTLLNVSL-TGIHADIQVR EAPRNADIQKGW
EPOR   -----ADTSSFFVPLELRVTAAS----GAPRYHRVHINEVLLDA VGLVARLADES-----GHVVLRL PPP--ETPMT
IL4R   -----DDVVSADNYTLDLWA----GQQLLWKGDFKPSSEHVK RA GNLTVHtnvs-----DTLLLT SNPY-PPDNLY
IL12R  DSACPAABESLPIEVMVDAVHK-LKYENYTSFFIRDIK D KNLQLKPLKN-----SRQVEVS EYPDWTSTPHSY
comBR  -----IQMA SLNVTKDG-----DSYSLR ETMKM---RYEH
          1F          1G          2A          2B

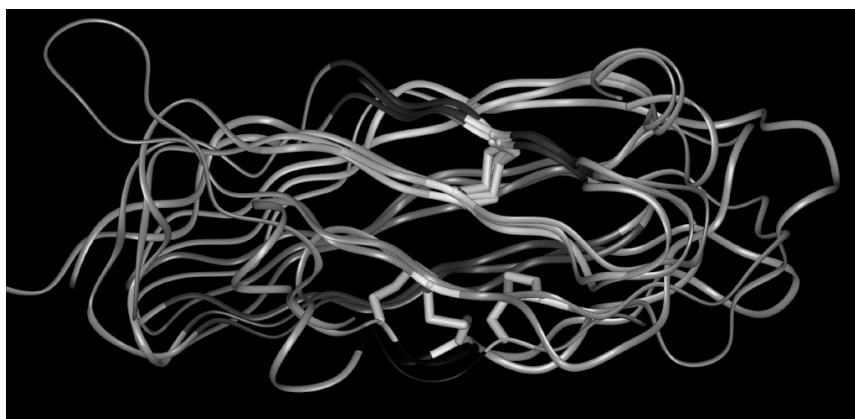
gp130  IILKYNIQVETKDA--STWSQIPPEDASTRSSFTVQDLKP--FTEYVFRIRCKED--GRGYWSDWSEASGITYEDR
mGCSFR MEQCEELRYQPQLKG-ANWTLVFL--PSSKDQFELCGLHQ--APVYTLQMRCHRSS--LPGFWSPWSPGLQLRPTM
rPLR   FTMEYETRLKPEEA--EEMEIFH-T--GHQTFKVFDFLYP--GQKYLVDTRCKPD---HGYSRWSQESSVEMP
GHR    MVLEVELQYREVNE--TKTKMMDPI----LTTSPVVSLLK--DKEYEVRVRSKORN---SGNYGEFSEVLYVTLF
EPOR   SHIRXEVVDSAGNGA-GSVQRVE-I--LEGRTECVLSNLRG--RTRYTFVARMAPSPFGEFSAWSEPVSLIT
IL4R   NHLTYAVNIWSENDP-ADFRIYV-V--TYLEPSLRIAastkGISYRARRVRAWAQA--YNTWSEWSPSTKWH
IL12R  FSLTFCVQVQKSKR-ekkdRVF-T--DKTSATVICRK-----NASISVRAQDRY--SSSWSEWASVPCS
comBR  IDHTFEQVETKDTATKDSKTET-LQ--NAHSMALPALEP--STRYWARVVRVTSRTGYNGIWSSEWSEARSWDTES
          2C          2C'          2E          2F          2G
    
```



Structural Alignment of Cytokine Receptors

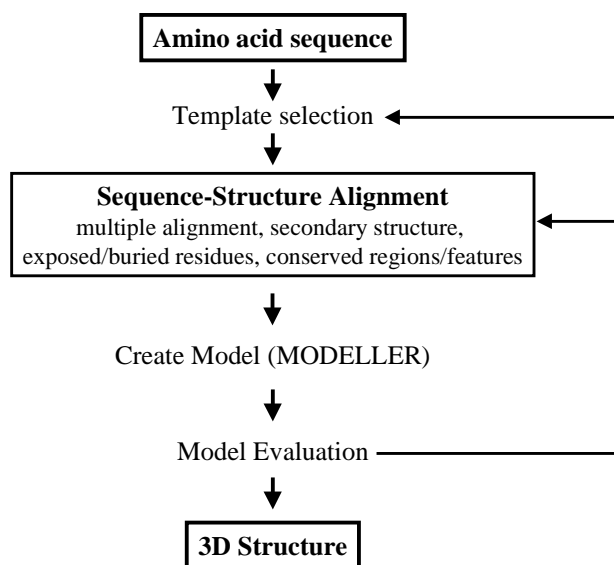


Things are not always structurally conserved !



LUDWIG
INSTITUTE
FOR
CANCER
RESEARCH

Comparative Modelling



LUDWIG
INSTITUTE
FOR
CANCER
RESEARCH

MODELLER

Sali, A. & Blundell, T.L. (1993) *J. Mol. Biol.* **234**, 779-815

- Determines spatial restraints from the template(s).
- Produces a number of 3D models of your sequence, best satisfying the template restraints.
- Conserved core residues are usually modelled well
- Loop regions are much more variable & difficult
- Loops are often functionally important

Be wary of web based modelling servers

If the sequence alignment is wrong, then so is the model!!

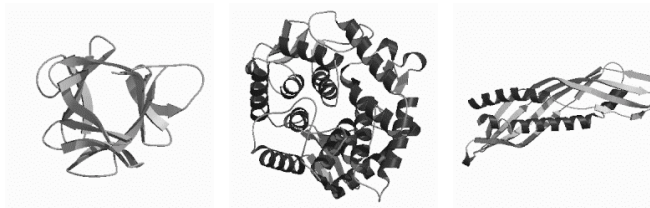


Threading/Tertiary Structure Prediction

- Predict structure of sequence with no known homology
- Align unknown sequence with structure (rather than sequence-sequence alignment)
- Use complex structure based energy scores rather than BLOSUM or PAM matrices

Useful for genome annotation (structure may infer function)

- Major Caveat - *novel folds, multiple domain proteins.*



Threading & Tertiary Structure Prediction Web Servers

genThreader

Jones, D. T. (1999) *J. Mol. Biol.* **287**, 797-815.

<http://insulin.brunel.ac.uk/psipred/>

fast threading algorithm

3D-PSSM

Kelley, L.A. et al, (2000) *J. Mol. Biol.* **299**, 499-520

<http://www.bmm.icnet.uk/~3dpssm/>

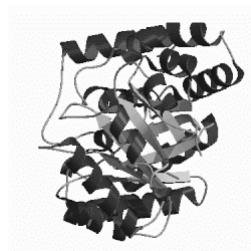
3D position specific scoring matrices



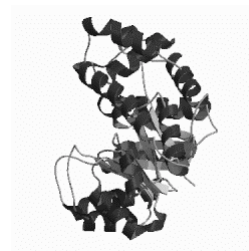
genTHREADER example

Confidence levels:

| | |
|--------|-------|
| CERT | > 99% |
| HIGH | 99% |
| MEDIUM | 90% |
| LOW | 70% |
| MARG | 40% |
| GUESS | < 40% |

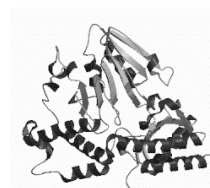


1dhp



1nal

1bt4



| | Conf | Prob | Epair | Esolv | AlnSc | Alen | DLen | Tlen | PDB_ID |
|---|-------|------|--------|-------|-------|------|------|------|--------|
| → | CERT | 1.00 | -76.2 | -17.4 | 864.0 | 291 | 292 | 300 | 1dhpA0 |
| → | CERT | 1.00 | -73.9 | -10.2 | 844.0 | 288 | 291 | 300 | 1nal30 |
| | CERT | 1.00 | -368.0 | -13.6 | 817.0 | 289 | 291 | 300 | 1nal10 |
| → | GUESS | 0.13 | -170.2 | 7.1 | 23.0 | 275 | 361 | 300 | 1bt4A0 |
| | GUESS | 0.10 | -165.5 | 3.6 | 22.0 | 189 | 213 | 300 | 1oroA0 |

Protein Docking

Protein-Protein Docking

- Many structures of individual proteins, far fewer complexes
 - Structural genomics projects
- Given the structure of two proteins, how do they interact?
 - what residues are involved in the interaction, what are the binding interfaces?
- Calculate interaction energies of many different orientations
- *Score* each orientation, predicting the best complex.
- Leads to hypothesis that can be tested - *not* reliable answers with current methods - need biological data



Protein Docking (2)

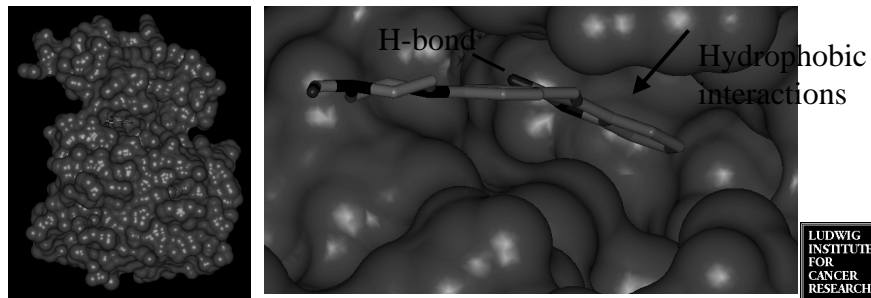
Protein-Ligand Docking

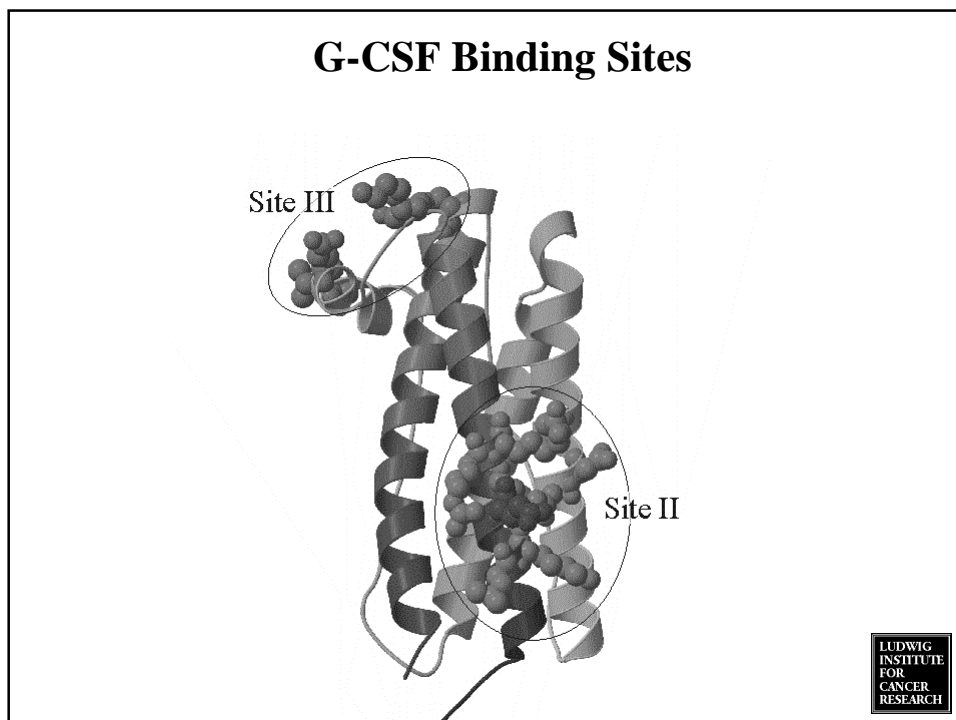
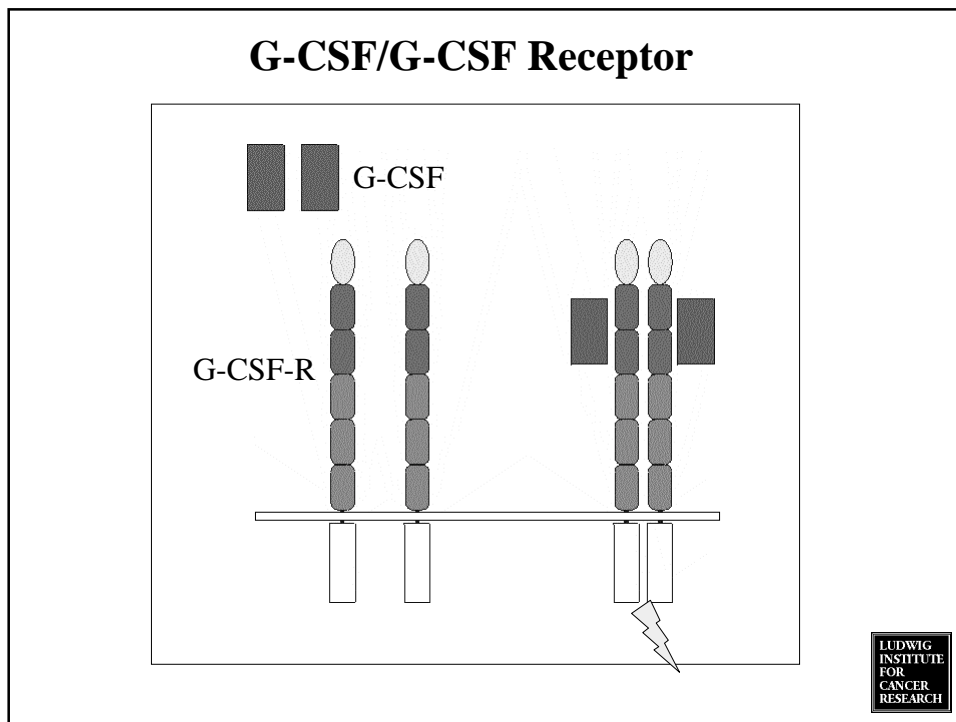
More advanced methods allow superior and more reliable results

- now being used for *in silico* screening

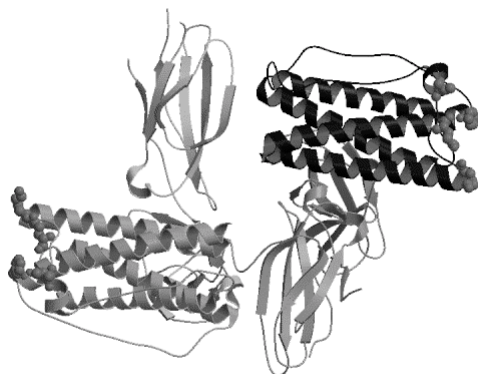
Considers H-bonds, electrostatics, hydrophobic interactions (burial)

Many biotech companies use these methods for drug discovery/development





Crystal G-CSF Receptor Complex



Aritomi, M. *et al.* (1999) *Nature* **401**, 713-717



gp130 viral IL-6 complex

Side View

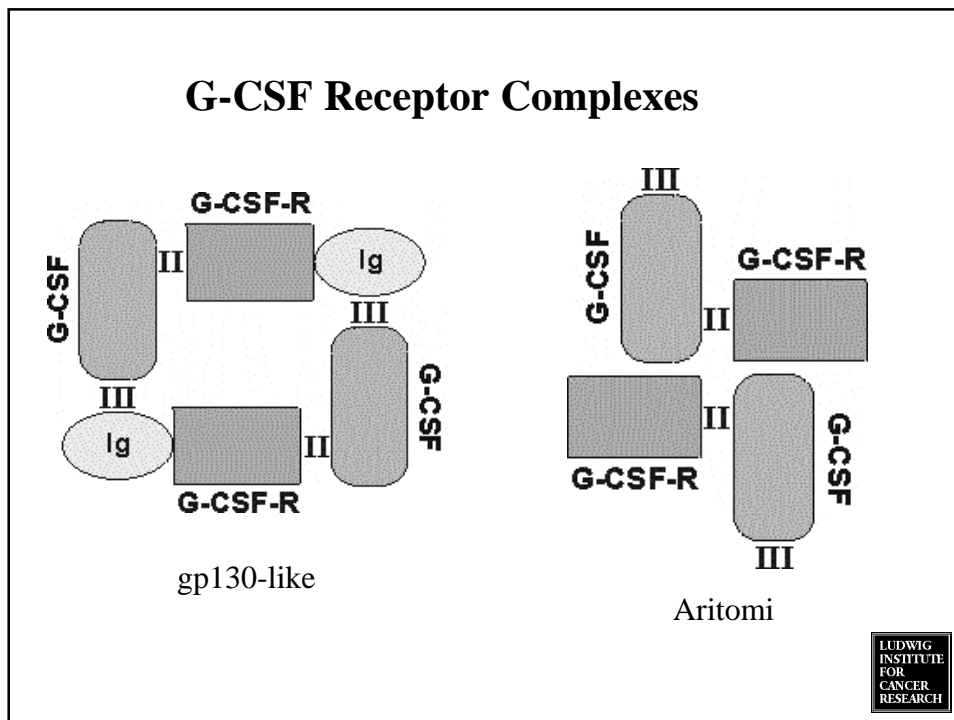


Top View



Chow, D *et al.* (2001) *Science* **291**, 2150-2155





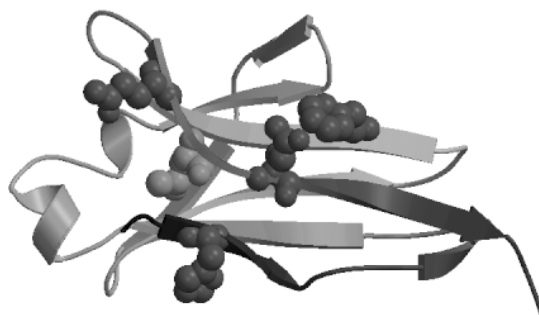
G-CSF Receptor Ig Domain

| | A | A' | B | α1 | C | C' | |
|---------------|--------|--------|-----------|---------|------------|------------------------------|---------------------|
| human GCSFR | EECS | CHISVS | APIVHLGDP | IT | ASCTIKQNC | SHLD--PEPQILWRL-GAELQ | |
| mouse GCSFR | ESCS | CHIEIS | PPVRLGDP | VLASCTI | SPNCSKLD-- | QQAKILWRLQDEPIQ | |
| cow GCSFR EST | EQCS | HILLSA | PVIRLGDQ | VITASCI | INRNC | SHLG--TDWRVVWKL-EPELH | |
| pig GCSFR EST | EECS | YISLPA | PIIRLGDPI | IVS | CIINQ | NCTSLG--PESQIRWKL-DTELQ | |
| mouse gp130 | QLLEPC | SYIY | PEFFV | VQRGS | NFTAT | CVLKEACLQHYVNASYIVWKT-NHAAV | |
| rat gp130 | QLVEPC | SYIY | PEFFV | VQRGS | NFTAT | CVLKEKCLQVYSVNATYIVWKT-NHVAV | |
| chick gp130 | GLVQSC | CHII | PESV | LALGS | NFTAL | CILNESCLDFGNIYASQIIWKM-KNKVI | |
| xenopus gp130 | ELVKVC | GRIF | PDGIV | HGER | PFTAY | CVINQTC | LR--DASRIYWLK-GKVKV |
| human gp130 | ELLDPC | SYIS | PESV | VQLHS | NFTAV | CVLKEKCMDFHVNANYIVWKT-NHFTI | |

| | D | E | F | G | G | | | | | | | |
|---------------|-------|----|----|---------|-------|-------------|------|--------------|-------|-------|--------------|---------|
| human GCSFR | EGGRQ | QR | LS | SDGTQES | II | TLPHLNHTQAF | LS | CCLNW--GNSL | QILDQ | VELRA | YYP | |
| mouse GCSFR | EGDRQ | HH | LP | DGTQES | LI | TLPHLNHTQAF | LS | CCLPW--EDSV | QLDQ | AE | LHAYYP | |
| cow GCSFR EST | PRERR | QH | LP | NGTLQ | STI | TLPHLNHSR | VLL | SCCLHW--GNSL | QILDQ | AE | LQAYYP | |
| pig GCSFR EST | EGGRQ | QH | LP | NGTLQ | STI | TFPHFNYS | RALL | SCCLHW--GNSL | QILDQ | AE | LQAYYP | |
| mouse gp130 | PREQV | T | V | INR-- | TTSSV | FTD | VVLP | SVQ | LT | CN | ILSFQIE | |
| rat gp130 | EKEQV | T | V | INR-- | TASSV | FTD | VVFN | VQ | LT | CN | ILSFQIE | |
| chick gp130 | EKEQY | R | E | INR-- | TVSSV | FTD | FN | SS | LA | S | PLTCNVLADQIE | |
| xenopus gp130 | RETQY | E | I | L | NQ-- | TTSSV | FTD | FN | LL | S | PLTCNVMAS | |
| human gp130 | EKEQY | T | I | INR-- | TASSV | FTD | IASL | NI | Q | LT | CN | ILTFQIE |

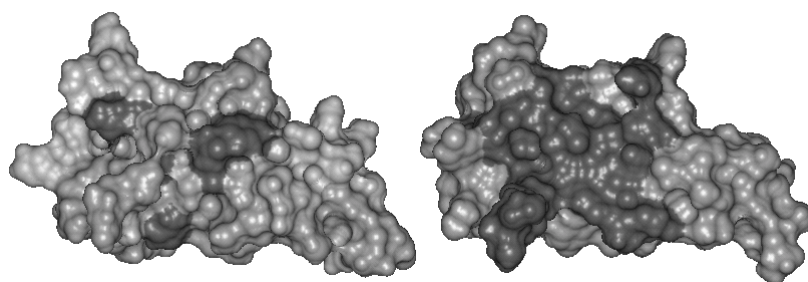
LUDWIG
INSTITUTE
FOR
CANCER
RESEARCH

G-CSF Receptor Ig domain Binding Residues



LUDWIG
INSTITUTE
FOR
CANCER
RESEARCH

Ig Domain Binding Surface Comparison



G-CSF-R

gp130

LUDWIG
INSTITUTE
FOR
CANCER
RESEARCH

