

The ENCODE project:

- *How many membrane proteins?*

Rita Casadio

BIOCOMPUTING GROUP
Interdepartmental Centre for Biotechnological
Research/Department of Biology
University of Bologna, Italy



Methods:

Prediction of the signal peptide

SPepLip

Method: NNs. Input: Single sequence

Fariselli P, Finocchiaro G, Casadio R (2003) *Bioinformatics* 19:2498-2499

Prediction of transmembrane α -helices

MEMSAT

Method: Dynamic programming. Input: Sequence profiles

Jones DT, Taylor WR, Thornton JM (1994) *Biochemistry* 15:3038-3049

TMHMM2.0

Method: HMMs. Input: Single sequence

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) *JMB* 305:567-580

ENSEMBLE 1.0+Filter *Method: NNs and HMMs. Input: Sequence profiles*

Martelli PL, Fariselli P, Casadio R (2003) *Bioinformatics* 19:I205-I211

TMHMMdomfix *Method HMMs. Input: Single profiles, SMART domains*

Bernsel A, von Heijne G (2005) *Prot Sci* 14:1723-1728

PRODIV_TMhMM *Method: HMMs. Input: Sequence profiles*

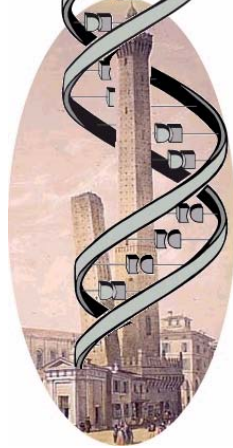
Viklund H, Elofsson A (2004) *Prot.Sci.* 13:1908-1917

Prediction of the bonding state of cysteines

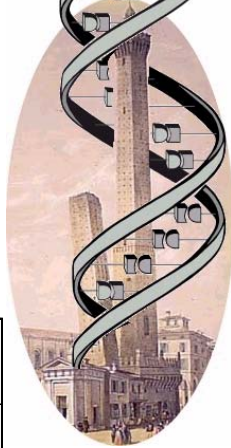
CYSPRED 2.0

Method: NNs and HMMs. Input: Sequence profiles

Martelli PL, Fariselli P, Casadio R (2004) *Proteomics* 4:1665-1671



Performance on the 121 high-resolved chains (from PDB)



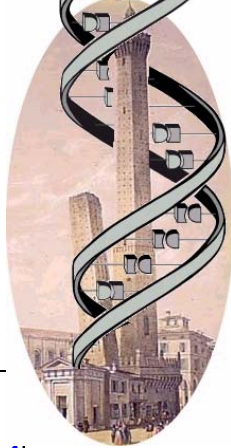
	$Q_{topography}$	$Q_{topology}$
TMHMM	88/121 (73%)	67/121 (55%)
TMHMMdomfix	87/121 (72%)	74/121 (61%)
→ PRODIV	99/121 (82%)	93/121 (77%)
MEMSAT	93/121 (77%)	90/121 (74%)
→ ENSEMBLE 1.0	105/121 (87%)	92/121 (76%)

Correct Topography: correct position of TMhelices along the sequence

Correct Topology: correct Topography AND correct Orientation of both N and C termini with respect to the membrane plane

Transmembrane proteins in the ENCODE data set

Whole set: 1,097 sequences

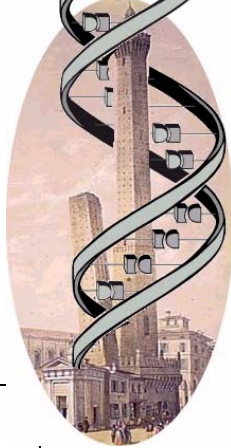


	Globular total	Globular with signal	TransMem total	TransMem with signal	TransMem with Bonded Cys
TMHMM	850 (77.5%)	130	247 (22.5%)	74	60
TMHMMdomfix	833 (75.9%)	130	264 (24.1%)	74	63
PRODIV	843 (76.8%)	129	254 (23.2%)	75	65
MEMSAT	702 (64.0%)	108	395 (36.0%)	96	107
ENSEMBLE 1.0 +Filter	822 (74.9%)	120	275 (25.1%)	84	71

Sequences predicted TM by at least one method: 424 (39%)
Sequences predicted TM by all the methods: 218 (21%)

Transmembrane proteins in the ENCODE data set

Only complete proteins: 748 sequences from 393 genes, 191 of which with 546 variants

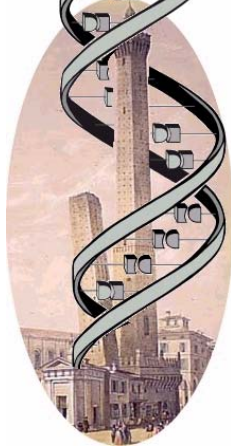


	Globular total	Globular with signal	TransMem total	TransMem with signal	TransMem with Bonded Cys
TMHMM	544 (72.7%)	102	204 (27.3%)	67	49
TMHMMdomfix	533 (71.3%)	102	215 (28.7%)	67	50
PRODIV	538 (71.9%)	99	210 (28.1%)	70	56
MEMSAT	439 (58.7%)	83	309 (41.3%)	86	88
ENSEMBLE 1.0 +Filter	523 (69.9%)	91	225 (30.1%)	78	61

Sequences predicted TM by at least one method: 327 (44%)
Sequences predicted TM by all the methods: 185 (25%)

Validation ?

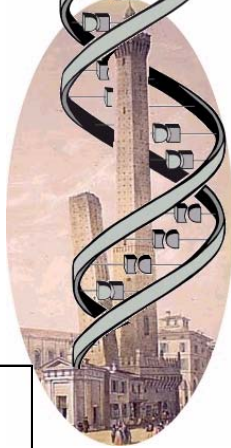
ENCODE proteins with a SwissProt Entry



Out of 313 chains, 72 contain Transmembrane in the Keyword field

	Q2	Q(TM)	Q(Glob)	P(TM)	P(Glob)	Corr
TMHMM	93.3%	79.2%	97.5%	90.5%	94.0%	0.81
TMHMMdomfix	92.3%	79.2%	96.3%	86.4%	93.9%	0.78
PRODIV	93.6%	80.6%	97.5%	90.6%	94.4%	0.82
MEMSAT	81.5%	83.3%	80.9%	56.6%	94.2%	0.57
ENSEMBLE 1.0+ Filter	93.4%	80.5%	97.1%	89.2%	94.4%	0.81

*Variants:
546 variants from 191 genes*

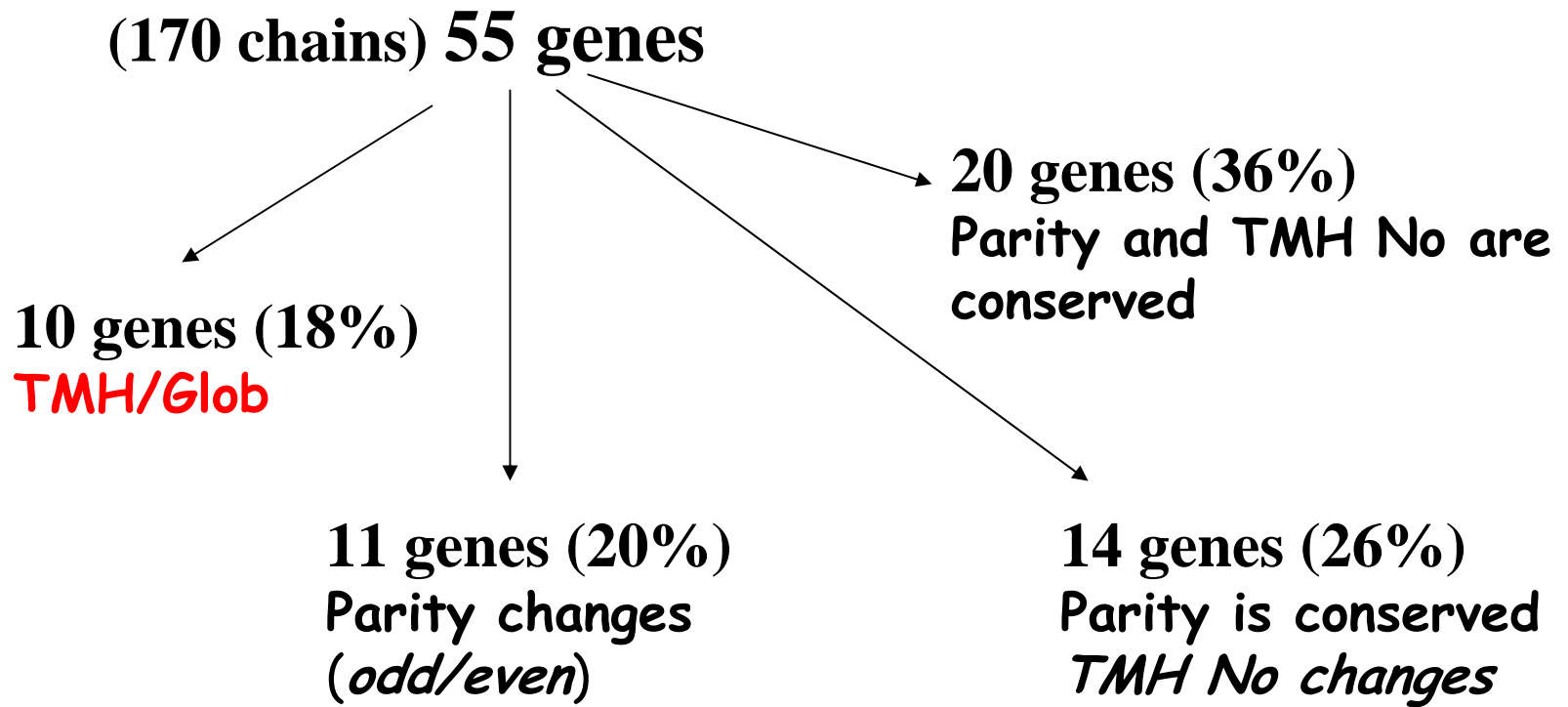


Genes that:

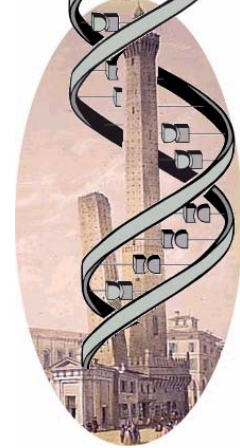
	Have at least a variant with TMH	Have variants with different number of TMHs	Have both TM and Glob variants
TMHMM	47 (140 seqs)	35	5
TMHMMdomfix	51 (151 seqs)	26	7
PRODIV	50 (149 seqs)	30	8
MEMSAT	82 (243 seqs)	48	20
ENSEMBLE 1.0+ Filter	55 (170 seqs)	35	10

Parity= Odd or Even No of TMHs

Prediction analysis of variants:



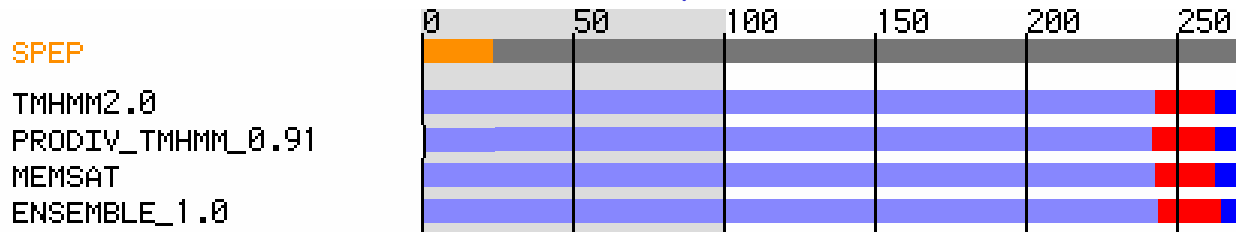
Variants- Example 1



AC0011501.5-001

Q5IHW5

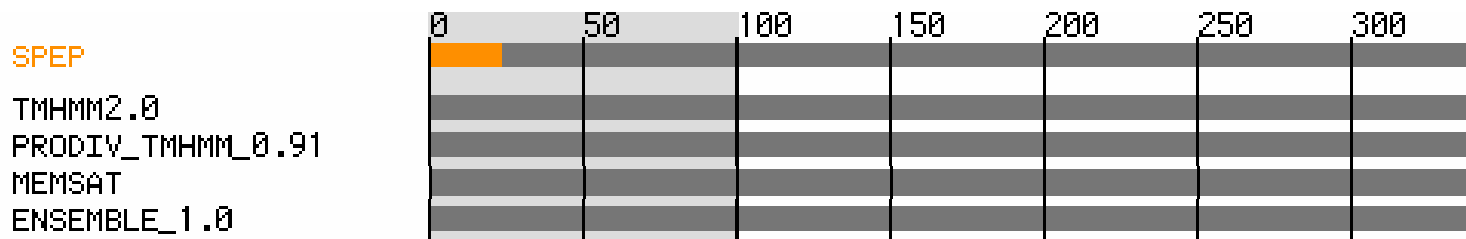
KIR2DL4



AC0011501.5-003

Q8N740

Killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 4 (Isoform 2) (KIR2DL4)

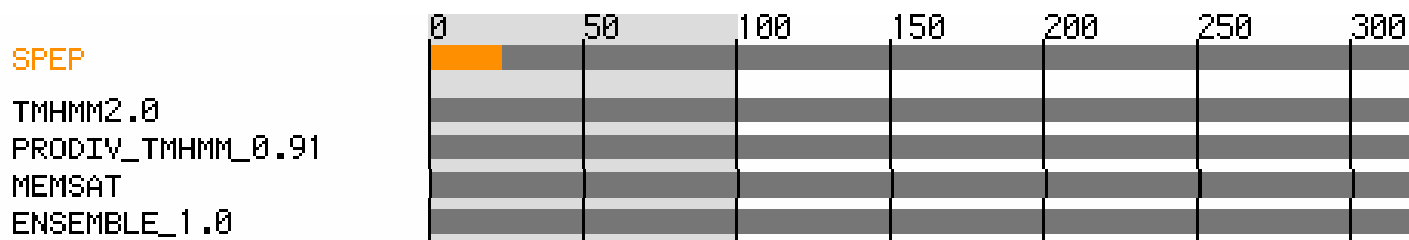


**Predicted
secreted**

AC0011501.5-004

Q8N739

Killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 4 (Isoform 6) (KIR2DL4)



**Predicted
secreted**

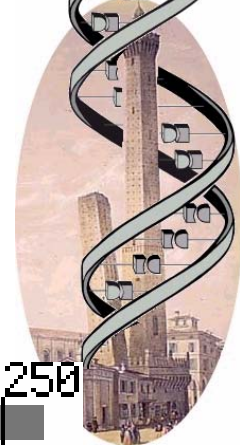
Legend
 SignalP: Signal Peptide
 inside: Cytoplasmic Loop
 outside: Extra Cytoplasmic Loop
 TMhelix: transmembrane helix
 glob: soluble (globular) protein
 Mature_protein: Mature protein

Variants

Example 1

AC011501.5-001	MSMSPTV IILACLGFFLDQSVWAHVGGQDKPFCSAWPSAVVPQGGHVTLRCHYRRGFNIF	60
AC011501.5-003	MSMSPTV IILACLGFFLDQSVWAHVGGQDKPFCSAWPSAVVPQGGHVTLRCHYRRGFNIF	60
AC011501.5-004	MSMSPTV IILACLGFFLDQSVWAHVGGQDKPFCSAWPSAVVPQGGHVTLRCHYRRGFNIF *****	60
AC011501.5-001	TLYKKDGVPVPELYNRIFWNSFLISPVTPAHAGTYRCRGFHPHSPTIEWSAPSNPLVIMVT	120
AC011501.5-003	TLYKKDGVPVPELYNRIFWNSFLISPVTPAHAGTYRCRGFHPHSPTIEWSAPSNPLVIMVT	120
AC011501.5-004	TLYKKDGVPVPELYNRIFWNSFLISPVTPAHAGTYRCRGFHPHSPTIEWSAPSNPLVIMVT *****	120
AC011501.5-001	GLYEKPSLTARPGPTVRAGENVTLS CSSQSSFDIYHLSREGEAHELRLPAVPSINGTFQA	180
AC011501.5-003	GLYEKPSLTARPGPTVRAGENVTLS CSSQSSFDIYHLSREGEAHELRLPAVPSINGTFQA	180
AC011501.5-004	GLYEKPSLTARPGPTVRAGENVTLS CSSQSSFDIYHLSREGEAHELRLPAVPSINGTFQA *****	180
AC011501.5-001	DFPLGPATHGETYRCFGSFHGSPYEWS D PSDPLPVSVT-----	218
AC011501.5-003	DFPLGPATHGETYRCFGSFHGSPYEWS D PSDPLPVSVTGNPSSSWPSPTEPSFKTDAAVM	240
AC011501.5-004	DFPLGPATHGETYRCFGSFHGSPYEWS D PSDPLPVSVT-----DAAVM *****	223
AC011501.5-001	-GNPS-----SSWSPTEPSFK-----TG-IARHL HAVIRYSVAILEFT	255
AC011501.5-003	NQEPAGHRTVNREDSDEQDPQEVTYAQLDHCIFTQRKITGPSQRSKRPSTDTSVCIELPN	300
AC011501.5-004	NQEPAGHRTVNREDSDEQDPQEVTYAQLDHCIFTQRKITGPSQRSKRPSTDTSVCIELPN *: * . * * :: ** * : . ** * * .	283
AC011501.5-001	ILPFLL HRWC SKKKMLL-----	273
AC011501.5-003	AEPRALSPAHEHHSQALMGSSRETTALSQTQLASSNVPAAGI	342
AC011501.5-004	AEPRALSPAHEHHSQALMGSSRETTALSQTQLASSNVPAAGI * * : . * :	325

Variants-Example 2

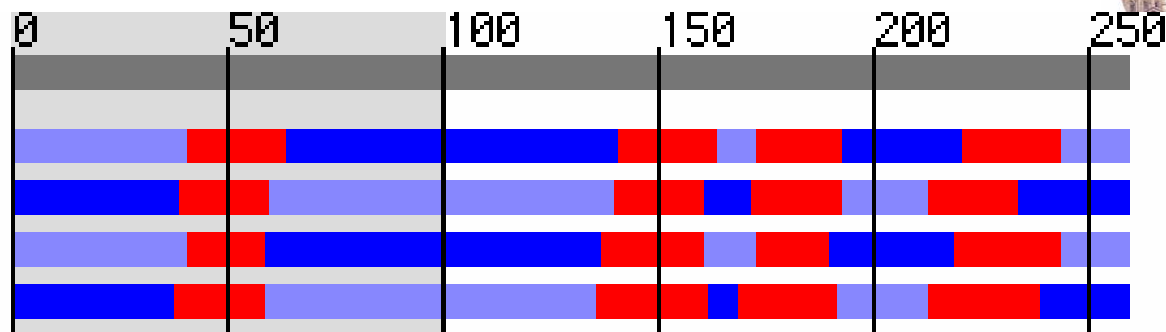


AC008440.9-001

CCG6_HUMAN

Neuronal voltage-gated calcium channel gamma-6 subunit (isoform a)

SPEP



TMHMM2.0

PRODIV_TMhMM_0.91

MEMSAT

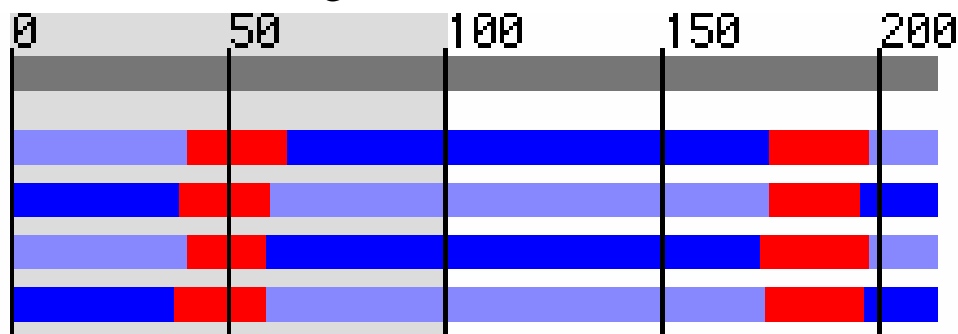
ENSEMBLE_1.0

AC008440.9-002

RefSeq:NM_145815

Neuronal voltage-gated calcium channel gamma-6 subunit (isoform b)

SPEP



TMHMM2.0

PRODIV_TMhMM_0.91

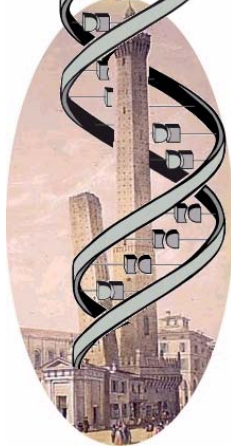
MEMSAT

ENSEMBLE_1.0

legend

SignalP: Signal Peptide
inside: Cytoplasmic Loop
outside: Extra Cytoplasmic Loop
TMhelix: transmembrane helix
glob: soluble (globular) protein
Mature_protein: Mature protein

Variants- Example 2



```
AC008440.9-001 MMWSNFFLQEENRRRGAAGRRAHGQGRSGLTPEREGKVKLALLLAAVGATLAVLSVGTE 60
AC008440.9-002 MMWSNFFLQEENRRRGAAGRRAHGQGRSGLTPEREGKVKLALLLAAVGATLAVLSVGTE 60
*****

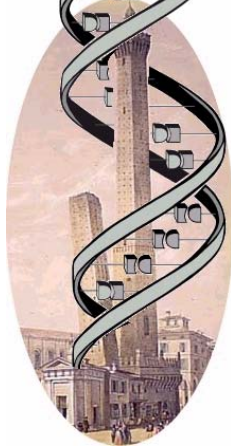
AC008440.9-001 FWVELNTYKANGSAVCEAAHLGLWKACTKRLWQADVPVDRDTCGPAELPGEANCTYFKFF 120
AC008440.9-002 FWVELNTYKANGSAVCEAAHLGLWKACTKRLWQADVPVDRDTCGPAELPGEANCTYFKFF 120
*****

AC008440.9-001 TTGENARIFQRTTKKEVNLAAVIAVLGLAVMALGCLCIIMVLSKGAEFLLRVGAVCFGL 180
AC008440.9-002 TTGENARIFQRTTKK----- 135
*****

AC008440.9-001 SGLLLLVSLEVFRHSVRALLQRVSPPEPPAPRLTYEYSWSLGCGVGAGLLILLGAGCFLL 240
AC008440.9-002 -GLLLLVSLEVFRHSVRALLQRVSPPEPPAPRLTYEYSWSLGCGVGAGLLILLGAGCFLL 194
*****

AC008440.9-001 LTLPSWPWGSLCPKRGHRAT 260
AC008440.9-002 LTLPSWPWGSLCPKRGHRAT 214
*****
```

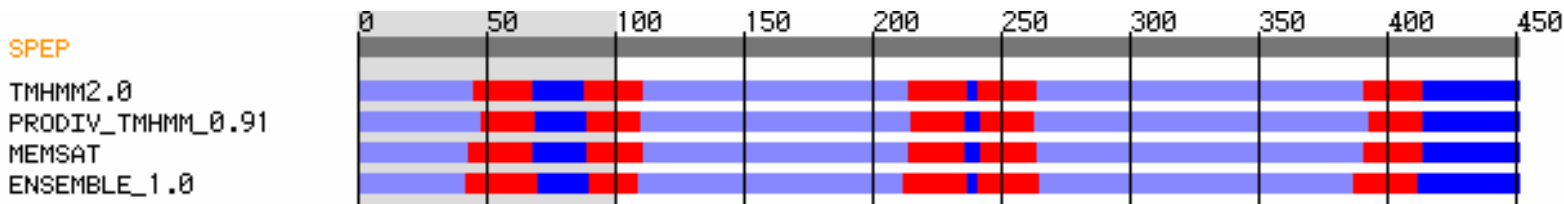
Variants - Example 3



AC008746.2-001

Q68A17

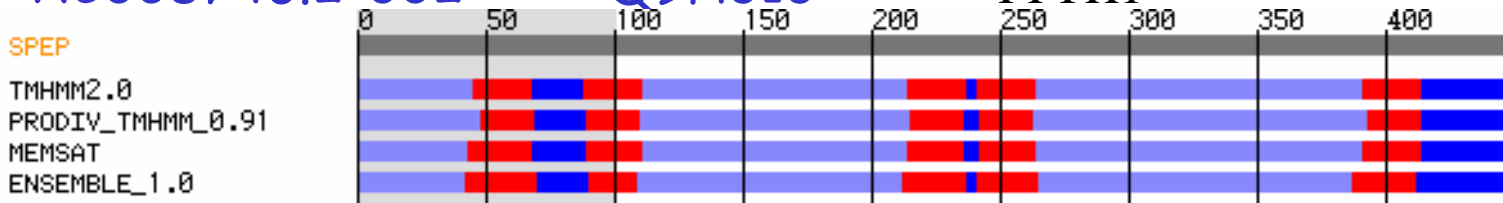
Tweety homolog 1



AC008746.2-002

Q9H313

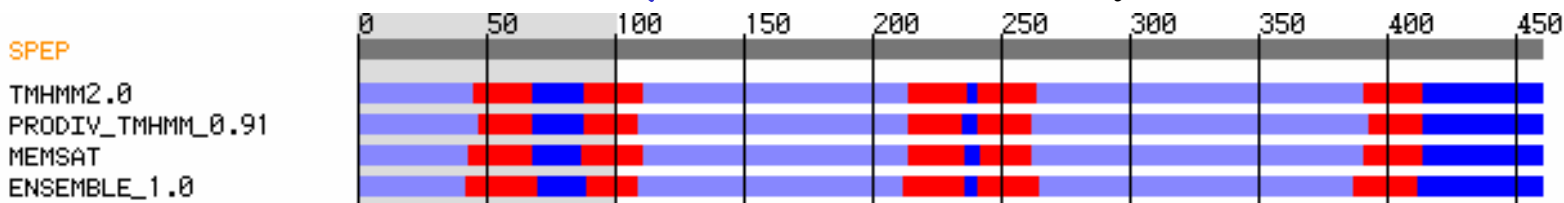
TTYH1



AC008746.2-003

Q5U682

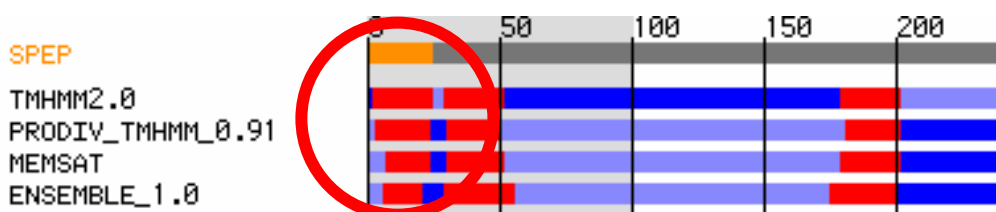
Tweety 1, isoform 2



AC008746.2-004

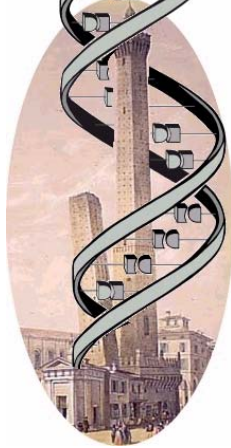
Q8WUU2

TTYH1 protein



legend

- SignalP: Signal Peptide
- inside: Cytoplasmic Loop
- outside: Extra Cytoplasmic Loop
- TMhelix: transmembrane helix
- glob: soluble (globular) protein
- Mature_protein: Mature protein



AC008746.2_001 MGAPPGYRPSAWVHLLHQLPRADFLQRPVPSVFAPQEQEYQQ**QALLLVAALAGLGLGLSLI** 60
AC008746.2_002 MGAPPGYRPSAWVHLLHQLPRADFLQRPVPSVFAPQEQEYQQ**QALLLVAALAGLGLGLSLI** 60
AC008746.2_003 MGAPPGYRPSAWVHLLHQLPRADFLQRPVPSVFAPQEQEYQQ**QALLLVAALAGLGLGLSLI** 60
AC008746.2_004 -----

AC008746.2_001 **FIAVYLIRF**CCCRPPEPPGSKIPSPGG**CVTWS**CIVALLAGCTGIGIGFYGNSETSDGVS 120
AC008746.2_002 **FIAVYLIRF**CCCRPPEPPGSKIPSPGG**CVTWS**CIVALLAGCTGIGIGFYGNSETSDGVS 120
AC008746.2_003 **FIAVYLIRF**CCCRPPEPPGSKIPSPGG**CVTWS**CIVALLAGCTGIGIGFYGNSETSDGVS 120
AC008746.2_004 -----

AC008746.2_001 QLSSALLHANHTLSTIDHLVLETVERLGEAVRTELTTLEEVLPRTELVA AARGARRQAE 180
AC008746.2_002 QLSSALLHANHTLSTIDHLVLETVERLGEAVRTELTTLEEVLPRTELVA AARGARRQAE 180
AC008746.2_003 QLSSALLHANHTLSTIDHLVLETVERLGEAVRTELTTLEEVLPRTELVA AARGARRQAE 180
AC008746.2_004 -----

AC008746.2_001 AAAQQLQGLAFWQGVPLSPLQVAENVSFVEEYR**WLAYVLLLLLELLVCLFTLLGLAKQSK** 240
AC008746.2_002 AAAQQLQGLAFWQGVPLSPLQVAENVSFVEEYR**WLAYVLLLLLELLVCLFTLLGLAKQSK** 240
AC008746.2_003 AAAQQLQGLAFWQGVPLSPLQVAENVSFVEEY**RWLAYVLLLLLELLVCLFTLLGLAKQSK** 240
AC008746.2_004 -----**MWLAYVLLLLLELLVCLFTLLGLAKQSK** 28

AC008746.2_001 **WLIVVMTVMSLLVLVLSWGS**MGLEAATAVGLSDFCSNPDPYVLNLTQEETGLSSDILSY 300
AC008746.2_002 **WLIVVMTVMSLLVLVLSWGS**MGLEAATAVGLSDFCSNPDPYVLNLTQEETGLSSDILSY 300
AC008746.2_003 **WLIVVMTVMSLLVLVLSWGS**MGLEAATAVGLSDFCSNPDPYVLNLTQEETGLSSDILSY 300
AC008746.2_004 **WLIVVMTVMSLLVLVLSWGS**MGLEAATAVGLSDFCSNPDPYVLNLTQEETGLSSDILSY 88

AC008746.2_001 LLCNRAVSNPFQQRLLTSLQRALANIHSQLLGLEREAVPQFP**SAQKPLLSLEETLN**VTEGN 360
AC008746.2_002 LLCNRAVSNPFQQRLLTSLQRALANIHSQLLGLEREAVPQFP**SAQKPLLSLEETLN**VTEGN 360
AC008746.2_003 LLCNRAVSNPFQQRLLTSLQRALANIHSQLLGLEREAVPQFP**SAQKPLLSLEETLN**VTEGN 360
AC008746.2_004 LLCNRAVSNPFQQRLLTSLQRALANIHSQLLGLEREAVPQFP**SAQKPLLSLEETLN**VTEGN 148

AC008746.2_001 FHQLVALLHCRSLHKDYGAALRGLCED**DALEGLLFLLLFSLLSAGALATALC**SLPRAWALF 420
AC008746.2_002 FHQLVALLHCRSLHKDYGAALRGLCED**DALEGLLFLLLFSLLSAGALATALC**SLPRAWALF 420
AC008746.2_003 FHQLVALLHCRSLHKDYGAALRGLCED**DALEGLLFLLLFSLLSAGALATALC**SLPRAWALF 420
AC008746.2_004 FHQLVALLHCRSLHKDYGAALRGLCED**DALEGLLFLLLFSLLSAGALATALC**SLPRAWALF 208

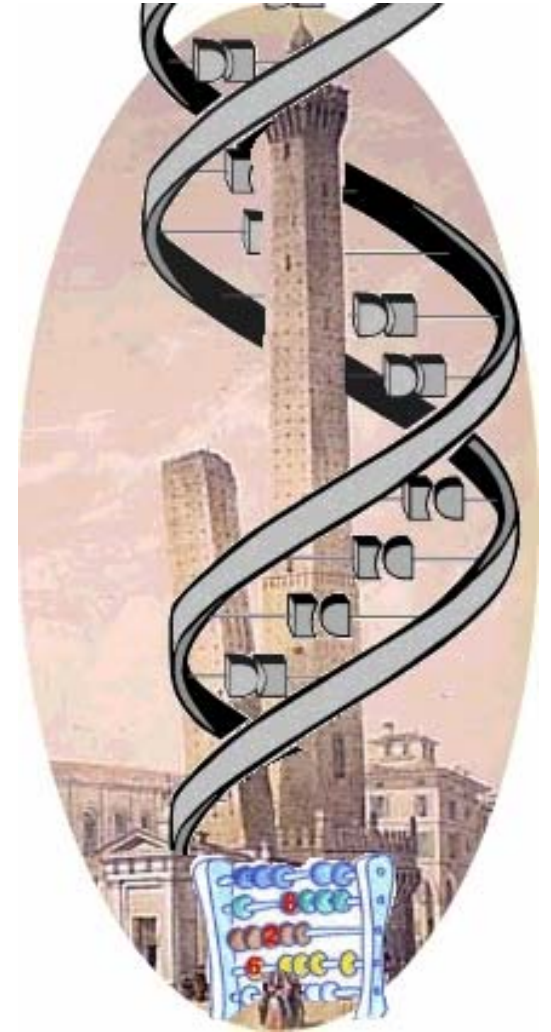
AC008746.2_001 PPSDD---YDDTDDDDP-----FNPQ**QESKRFVQWQSSI-** 451
AC008746.2_002 PPSDD---YDDTDDDDP-----FNPQ**-ESKRFVQWQSSI-** 450
AC008746.2_003 PPRNPSALCSGSRLLSEPLLPAGLEPGSPLRSFPGCRRRPH 460
AC008746.2_004 PPSDD---YDDTDDDDP-----FNPQ**QESKRFVQWQSSI-** 239

Variants Example 3

** : ..: .:* ::* : * :

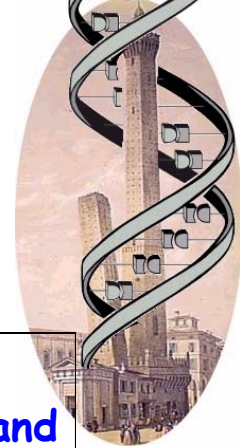
The ENCODE project:

- *How many proteins with S-S bridges?*



The Biocomputing Group
University of Bologna

Cysteine bonding state in the ENCODE data set



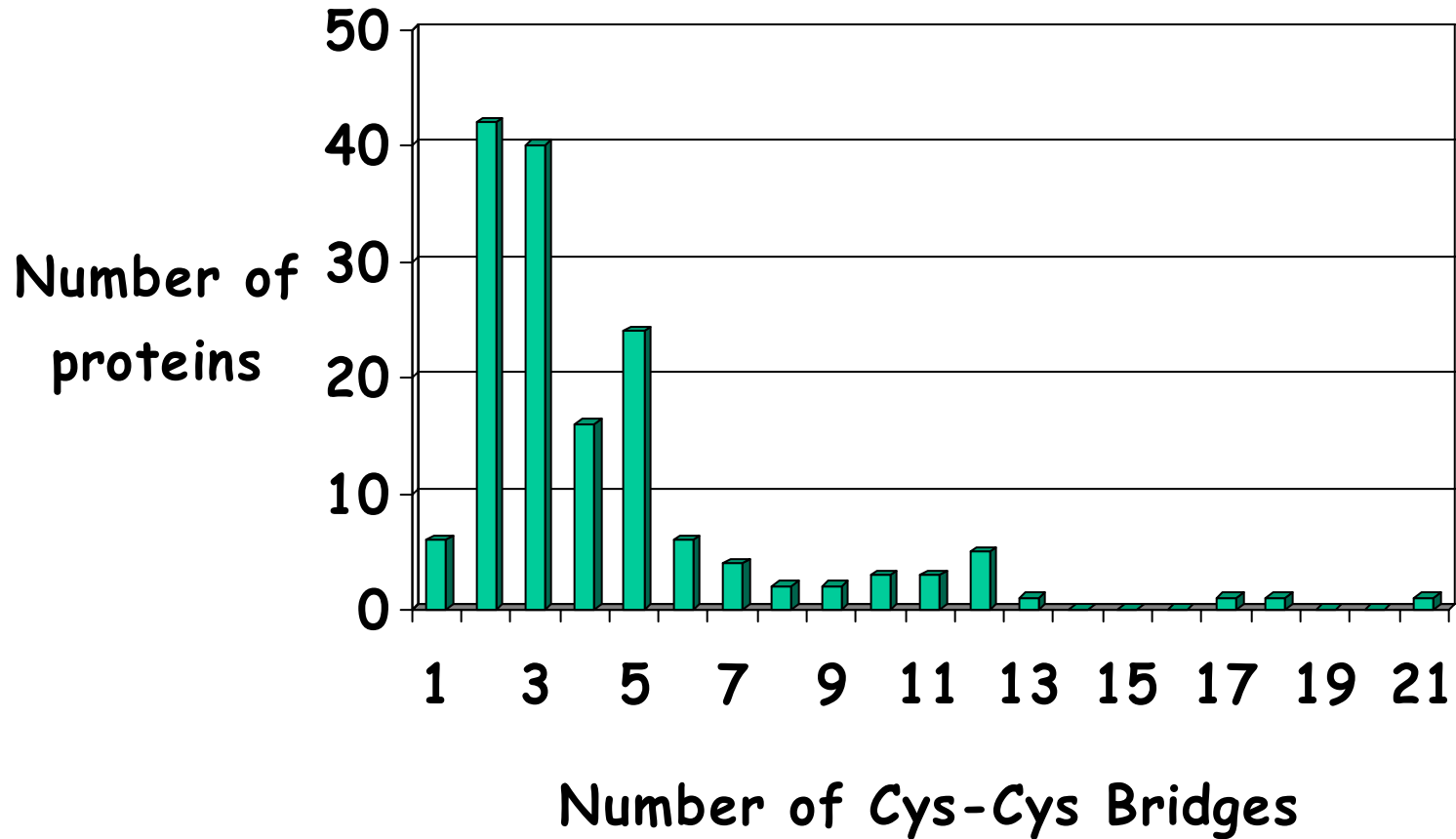
Whole set: 1,097 sequences

Number of Bonded Cys	Number of free Cys	Proteins with Bonded Cys	Proteins with Free Cys	Proteins with both Bonded and Free Cys
1,940 (20.0%)	7,749 (80.0%)	205 (18.7%)	923 (84.1%)	108 (9.8%)

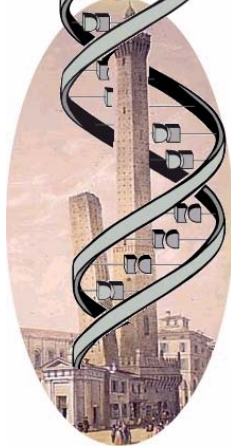
Only complete proteins: 748 sequences

Number of Bonded Cys	Number of free Cys	Proteins with Bonded Cys	Proteins with Free Cys	Proteins with both Bonded and Free Cys
1,476 (20.4%)	5,764 (79.6%)	159 (21.3%)	634 (84.8%)	81 (10.8%)

Cysteine bonding state in the ENCODE data set
Only complete proteins: 748 sequences



Validation



Scoring Indexes

Overall Accuracy

$$Q2 = \#(\text{correct predictions}) / \#(\text{examples})$$

Per class accuracy

$$Q(x) = \#(\text{correct prediction in class } x) / \#(\text{examples in the class } x)$$

Probability of correct prediction (per class)

$$P(x) = \#(\text{correct prediction in class } x) / \#(\text{predictions in the class } x)$$

Correlation

Corr: Matthews correlation index

The Biocomputing Group of the University of Bologna

