Insilico Functional Annotation of Hypothetical ORFs in Human Chromosome2

Sivashankari Selvarajan[1] and Piramanayagam Shanmughavel [2]
[1]*Assistant Professor in UGC – Innovative Programme Department of Bioinformatics*
Nirmala College for Women, India
Email: sivpongiannan@gmail.com
[2]*Assistant Professor, Department of Bioinformatics, Bharathiar University, India*
Email: shanvel_99@yahoo.com

Abstract

The high-throughput genome projects have resulted in a rapid accumulation of genome sequences for a large number of organisms and large number of genes with unknown function (Hypothetical). To fully realize the value of the data, scientists need to identify proteins encoded by these genes and understand how these proteins function in making up a living cell. With experimentally verified information on protein function lagging far behind, computational methods are needed for reliable and large-scale functional annotation of proteins. Functional annotation is the process of identifying for a given gene its biological function, interaction with other elements, involvement in metabolic pathways, and any other piece of information that helps in understanding when and how a gene influences the overall system. On the other hand, many Biological Processes and Disease mechanisms are still unknown due to lack of knowledge about the function of the Hypothetical genes in Human. Once its function is revealed the so called hurdle of unknown mechanism of the Human Genome can be mastered. Hence, the present study aims to use computational approaches to annotate the function of hypothetical genes in Chromosome 2 of Human. The annotation of the hypothetical genes in human chromosoem2 was done both at the nucleotide and protein level. Among the 41 uncharacterized hypothetical genes in Human chromosome 2, the functions of 27 of them were successfully annotation. Further, experimental validation is essential to confirm the predicted function.

Keywords: *Annotation, Chromosome2, Function, Hypothetical genes*

Background

The high-throughput genome projects have resulted in a rapid accumulation of genome sequences for a large number of organisms and large number of genes with unknown function (Hypothetical). In biochemistry, a hypothetical protein encoded by a hypothetical gene is a protein whose existence has been predicted, for which there is no experimental evidence for expression *in vivo (*Zarembinski *et al* 1998). As a result, the function of such genes is not known. This is due to the fact that they are predicted using computational methods, which rely on signals in DNA sequences to predict it as a gene or based on similarity to genes in other organisms. In this case, the function of these homologous genes is also not known. Not only in Human Genome, in all genomes sequenced to date, a large portion of these organisms‟ protein coding regions encodes polypeptides of unknown biochemical, biophysical, and/or cellular functions.

The usual scenario involving a hypothetical protein is in gene identification during genome analysis. When the bioinformatics tool used for the gene identification finds a large open reading frame without an analog in the protein database, it returns "hypothetical protein" as an annotation remark. To fully realize the value of the data, scientists need to identify proteins encoded by these genes and understand how these proteins function in making up a living cell.

Despite several efforts, only 50-60 % of genes have been annotated in most completely sequenced genomes and their functions are known. The rest 40% of the genes in any genome

is totally unknown in terms of its functions. The experimental characterization of such a huge number of hypothetical genes will take many decades before the biological function encoded by such hypothetical genes is known.

As of September 2014, there are around 637 genes encoded as Hypothetical in NCBI. These hypothetical ORFs may be functionally important and play very important roles in growth, development and maintenance of *Homo sapiens*. Research is needed to unravel the function of these conserved hypothetical genes in Human to understand more about molecular mechanisms and biological significance of the entire Human Genome. The 637 hypothetical ORFs in the Human Genome are encoded as „Hypothetical" because its expression and existence is not proved and hence its function is also unknown.

Many Biological Processes and Disease mechanisms are still unknown due to lack of knowledge about the function of these Hypothetical genes. Once its function is revealed the so called hurdle of unknown mechanism of the Human Genome can be mastered. Automated genome sequence analysis and annotation may provide ways to understand genomes. Thus, determination of protein function is one of the challenging problems of the post-genome era. This demands bioinformatics to predict functions of un-annotated protein sequences by developing efficient tools and methods. In addition, previous studies on hypothetical genes in other organisms have revealed that many hypothetical proteins are expressed and are involved in many important biological functions (Tobias *et al*., 2008; Chen *et al*., 2003). Similarly, experimental analysis on human embryonic stem lines reports proof for expression of hypothetical genes in the Human Genome (Bhattacharya *et al.*, 2004). On the other hand, a hypothetical gene *FLJ30473* in chromosome 22 was found to express in multiple human tissues, including brain, colon, heart, kidney, liver, lung, muscle, ovary, pancreas, placenta, small intestine, and testis and is homologous to apoptosis-inducing factor (Xie *et al*., 2005). But they employ either computational and experimental methods or just experimental methods only. However, computational functional annotation of Human proteins revealed functional descriptors for 7 Hypothetical genes (Roy, 2008). But high throughput functional annotation of hypothetical genes in the Human Genome using computational methods is not yet attempted and experimental methods are laborious and time consuming. So, the present investigation focused on functional annotation of hypothetical genes in the Human Genome with the following Objectives:

1. To identify the uncharacterized hypothetical Genes in chromosome2 of the Human Genome.
2. To annotate the function of the uncharacterized hypothetical genes in the human Genome both at gene and Protein level.
3. To assign functional categories for the annotated hypothetical genes.
4. To assign transmembrane Topology and Sub-Cellular Localization for the unannotated Hypothetical Genes.

Methodology

The Hypothetical ORFs in chromosome 2 of the Human Genome was retrieved from NCBI [Geer *et al*., 2010]. To identify whether the ORFs can actually be genes two strategies were followed. Initially, conservation of the ORF in other organisms was determined using Homologene[Geer *et al*., 2010] and then its coding potential was calculated using Coding Potential Calculator (CPC) [Lei Kong et al., 2007]. Conservation and coding potential were determined, because the ORFs have a high probability to be functional if they are conserved and having a higher coding potential score.

Next, the annotation of the hypothetical genes which are conserved and with coding potential at the nucleotide level was done using BLAST2GO [Ana *et al*., 2005] with the following algorithm: Initially, the sequence was queried against BLAST to find homologs followed by mapping of the sequence with GO terms including running Interproscan. At the protein level, pfam [Robert *et al*., 2013] and supfam database[Pandit *et al*., 2002] were used to assign domains and superfamily to the hypothetical proteins. Finally, COG[Geer *et al*.,

2010] and SCOP[Murzin et al., 2002] were used to assign functional category to the hypothetical genes.

The transmembrane topology and subcellular localization were predicted for the unannotated hypothetical genes using TMHMM[Krogh et al., 2001] and PSort[Nakai and Horton, 1999]

Results and discussion

The 41 hypothetical ORFs present in chromosome 2 of the human genome were retrieved from NCBI and it was found that, it contains 9 characterized genes. The remaining 32 hypothetical ORFS were tested to find whether they can be functional and non-functional. For this purpose, Homologene and Coding Potential Calculator were used. The results are tabulated in Table 1.

Table 1 Conservation and Coding Potential of Hypothetical Genes in Human Chromosome 2

Sl No.	**Gene Name**	**Conservation**	**Coding Potential**
1	C2orf27A	*Pan troglodytes*	3.69038
2	C2orf88	*Pan troglodytes, Bos taurus, and Mus musculus.*	1.34178
3	C2orf61	*Pan troglodytes, Mus musculus*	3.71202
4	C2orf65	*Pan troglodytes, Canis lupus familiaris, Bos taurus, Mus musculus, Gallus gallus, and Danio rerio.*	8.58827
5	C2orf54	*Pan troglodytes, Bos taurus, Mus musculus, Rattus norvegicus, and Gallus gallus.*	11.9126
6	C2orf63	*Pan troglodytes, Canis lupus familiaris, Bos taurus, Mus musculus, and Rattus norvegicus*	10.1827
7	C2orf74	*Pan troglodytes, Canis lupus familiaris, Mus musculus, and Rattus norvegicus.*	1.99556
8	C2orf57	*Pan troglodytes, Canis lupus familiaris, Bos taurus, Mus musculus, and Rattus norvegicus*	7.0881
9	C2orf82	*Bos taurus, Mus musculus, Rattus norvegicus, and Gallus gallus.*	1.858
10	C2orf80	*Pan troglodytes, Canis lupus familiaris, Bos taurus, Mus musculus, and Danio rerio.*	5.41768
11	C2orf72	*Pan troglodytes, Bos taurus, and Mus musculus.*	5.57153
12	C2orf67	*Pan troglodytes, Canis lupus familiaris, Bos taurus, Musmusculus, Rattus norvegicus, Gallus gallus,* and Danio rerio	10.6203
13	C2orf68	*Canis lupus familiaris, Bos taurus, Mus musculus, Rattus norvegicus, and Danio rerio*	1.74447
14	C2orf81	*Pan troglodytes, Mus musculus, and Rattus norvegicus*	11.8126
15	C2orf29	*Pan troglodytes, Canis lupus familiaris, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, Danio rerio, Drosophila melanogaster, Anopheles gambiae, Arabidopsis thaliana, and Oryza sativa*	8.08018
16	C2orf50	*Pan troglodytes, Canis lupus familiaris, and Rattus norvegicus*	3.16047
17	C2orf53	*Bos Taurus*	6.01493
18	C2orf43	*Pan troglodytes, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, Danio rerio, Drosophila melanogaster, Anopheles gambiae, Caenorhabditis elegans, Arabidopsis thaliana, and Oryza sativa.*	4.38865
19	C2orf76	*Pan troglodytes, Canis lupus familiaris, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, and Danio rerio.*	2.37845

20	C2orf47	*Canis lupus familiaris, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, and Danio rerio.*	5.27806
21	C2orf77	*Pan troglodytes, Canis lupus familiaris, Bos taurus, Mus musculus, Rattus norvegicus, and Gallus gallus.*	12.1034
22	C2orf73	*Pan troglodytes, Canis lupus familiaris, Bos taurus, and Rattus norvegicus*	5.9361
23	C2orf70	*Pan troglodytes, Canis lupus familiaris, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, and Danio rerio.*	5.25668
24	C2orf69	*Pan troglodytes, Canis lupus familiaris, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, and Danio rerio.*	7.22082
25	C2orf84	*Bos taurus, Mus musculus, and Rattus norvegicus*	3.19605
26	C2orf66	*Canis lupus familiaris, Bos taurus, Mus musculus, Rattus norvegicus, and Gallus gallus.*	1.36366
27	C2orf62	*Pan troglodytes, Canis lupus familiaris, Bos taurus, Mus musculus, Rattus norvegicus, and Danio rerio*	9.45344
28	C2orf85	*Pan troglodytes, Canis lupus familiaris, Bos taurus*	9.12913
29	C2orf15	*Pan troglodytes*	0.967256 *
30	C2orf27B	*Pan troglodytes*	4.77747
31	C2orf48	*Not Conserved*	0.774314 *
32	C2orf16	*Canis lupus familiaris, Bos Taurus*	8.16544

*Weak Coding Potential

From Table 1, it is clear that, all the hypothetical genes except C2orf48 in Chromosome 2 of human are conserved; 30 genes (73%) are conserved in *Mus musculus*, 32 genes (78%) are conserved in *Pan troglodytes*, 30 genes are conserved in *Bos Taurus* (73%), 27 genes (65%) are conserved in *Rattus norvegius*, 27 genes (65%) are conserved in *Canis lupus familiaris*, 18 genes (43%) are conserved in *Gallus gallus*, 17 genes (41%) are conserved in *Danio rerio*, 5 genes are conserved in *Drosophila melanogaster*, 5 genes are conserved in *Anopheles gambiae*, 2 genes are conserved in *Caenorhabditis elegans*, 1 gene in *Plasmodium falciparum*, 3 genes are conserved in *Arabidopsis thaliana and Oryza sativa.* All the hypothetical genes in Chromosome2 of human have strong coding potential for proteins except C2orf15 and C2orf48 which have weak coding potential.

The Superfamily identification using Superfamily database revealed the domains of 4 hypothetical genes and interproscan in BLAST2GO identified another 5 important domains among the 32 uncharacterized hypothetical genes in the human genome. Among the predicted domains, the hypothetical genes „C2orf82" contain a conserved domain called „AF0104" whose function is not annotated according to SCOP category. Similarly the „ARM REPEAT" identified within „C2orf63" by SUPFAM did not have functional annotation but the interproscan has successfully identified glycoside hydrolase within this hypothetical protein. This is due to the fact that failure of one annotation method can be compensated by usage of multiple methods resulting in identification of the function of hypothetical genes. Since, earlier studies have revealed that, hypothetical proteins contain fewer Pfam domains than known genes and the majority of these domains found in hypothetical proteins are annotated as "Domains of Unknown Functions (DUFs)" (Ramachandran *et al.*, 2009), three more Uncharacterized domains were identified by pfam.

The BLAST2GO annotation has identified the two hypothetical genes „C2orf27a" and „C2orf62" to encode a protein with kinase activity and two hypothetical genes („C2orf29" and „C2orf74") with oxidoreductase activity. The another two hypothetical genes C2orf48 and C2orf16 has metal binding activity and phosphatase activity respectively, accounting for a total of 6 hypothetical genes to involve in important metabolic process of the human genome. The results are presented in Table 2. Soumelis *et al*., (2010) has discussed about the presence of more number of proteins and enzymes in chromosome 2 for information processes such as replication, transcription and translation because of the speed in which

these process takes place. In accordance with this study two more genes „C2orf15" and „C2orf48" are involved in transcription among the hypothetical genes in chromosome 2. All these hypothetical genes were assigned functional category using SCOP or COG. The functional categories of these hypothetical genes using SCOP and COG are presented in Table 3.

Table 2 Functional annotation of Hypothetical Genes in Chromosome 2

Sl No.	GENE NAME	COMPONENT, PROCESS, FUNCTION	DOMAIN / SUPER FAMILY	FUNCTION	FUNCTIONAL CATEGORY	
					COG	SCOP
1	C2ORF27A	C: Endomembrane System; C: Trans-Golgi Network; C: Cytosol; P: Protein Localization; F: Protein Kinase Binding; C: Plasma Membrane	-	Regulation: Signal Transduction	T	-
2	C2ORF88	C: Extracellular Region	Copper Type II, Ascorbate-Dependent Monooxygenase	Metabolism: Oxidation/ Reduction	-	RA
3	C2ORF63	F: Binding; P: Carbohydrate Metabolic Process	Glycoside Hydrolase; ARM Repeat	Metabolism: Other Enzymes	-	RC;R
4	C2ORF74	C: Integral to Membrane F:Oxidoreductase Activity	Short-Chain Dehydrogenase/Reductase	Metabolism: Oxidation/Reduction	-	RA
5	C2ORF82	C: Integral to Membrane	AF0104/ALDC/Ptd012-Like	Not Annotated	-	NONA
6	C2ORF72	F: Protein Binding C: Extracellular Region	Heat Shock Protein 70	Cellular Processes: Posttranslational Modification, Protein Turnover, Chaperones	O	-
7	C2ORF29	F: Molecular P: Cell Proliferation C: Cellular Component P: Cell Proliferation; F: Zinc Ion Binding; F: Oxidoreductase Activity; P: Oxidation Reduction	Alcohol Dehydrogenase, Zinc-Containing	Metabolism: Energy Production and Conversion	C	-
8	C2ORF43	F: Catalytic Activity	Alpha/Beta-Hydrolases	Metabolism: Other Enzymes	-	RC
9	C2ORF47	C: Mitochondrion C: Nucleus; F: Protein	-	General	-	R

		Binding; C: Cytoplasm				
10	C2ORF77	-	Tropomyosin	Processes_IC : Cellmotility	-	N
11	C2ORF62	F: Molecular Function; P: Biological Process; P: Signal Transduction; F: CAMP-Dependent Protein Kinase Regulator Activity	-	Cellular Processes: Signal Transduction Mechanisms	T	-
12	C2ORF85	C: Integral to Membrane	-	General	-	R
13	C2ORF15	F: DNA Binding; C: Membrane; C: Nucleus; F: Zinc Ion Binding; P: Regulation Of Transcription, DNA-Dependent	-	Information Storage and Processing: Transcription	K	
14	C2ORF27B	C: Integral to Membrane	-	General	-	R
15	C2ORF48	C: Intracellular Membrane-Bounded Organelle; F: Metal Ion Binding; C: Cytoplasmic Part; P: Transcription; P: Developmental Process; C: Membrane	-	Information storage and processing: Transcription	K	-
16	C2ORF16	P: Protein Amino Acid Dephosphorylation; F: Protein Tyrosine Phosphatase Activity	-	Regulation: Kinases and Phosphatases and Inhibitors	-	OB
17	C2ORF61	-	Cag-Z	Other: Unknown Function	-	S
18	C2ORF70	-	Galactose-Binding Domain-Like	Metabolism: Carbohydrate Transport and Metabolism	-	G

Table 3 Functional Categories for the Hypothetical Genes in Chromosome 2

S.No	**Group Code**	**Description**	***Number a***
COG category			
1	J, A, K, L, B	Information storage and processing	*2*

2	D, Y, V, T, M, N, Z, W, U, O	Cellular processes and signaling	*3*
3	C, G, E, F, H, I, P, Q	Metabolism	*1*
4	R, S	Poorly characterized	-
SCOP Category			
5	RF, RE, P, MA, RG, SB, D, IA, N, NA, O, OA	PROCESSES	*1*
7	CA, C, CB, E, EA,F, G, GA, GB, H, RA, RB, RC, I, M, Q	METABOLISM	*5*
8	HA, HB, HC, HE, R, RD, ST	GENERAL	*4*
9	B, J, K, L,LB, Y	INFORMATION	-
10	A, LA, OB, T, TA, HD	REGULATION	*1*
10	S, SA	OTHER	*1*
11	*NONA*	*NOT ANNOTATED*	*1*

a „–"indicates there is no newly annotated gene in this COG or SCOP functional category.

From table 3, it is evident that maximum number of hypothetical genes in chromosome2 are involved in metabolism according to SCOP functional category, which coheres with the study of Antony *et al* (1999).

Thus, among the 32 uncharacterized hypothetical genes in chromosome 2 of the human genome, functional annotation and functional categories were assigned successfully to 14 of them. In spite of serious efforts functional assignment for 14 of them could not be done, however their sub cellular localization and possibility of presence in membrane are presented in Table 4.

Table 4 Topology and Localization of Un-annotated Hypothetical Genes in Chromosome 2

S.No.	**Gene**	**Subcellular Localization**	**TM Topology**
1	C2orf65	Nucleus	No
2	C2orf54	Plastid	No
3	C2orf57	Nucleus	No
4	C2orf80	Nucleus	No
5	C2orf67	Nucleus	No
6	C2orf68	Nucleus	No
7	C2orf81	Nucleus	No
8	C2orf50	Nucleus	No
9	C2orf73	Cytoplasm	No
10	C2orf69	Mitochondria	No
11	C2orf84	Cytoplasm	No
12	C2orf53	Nucleus	No
13	C2orf76	Cytoplasm	No
14	C2orf66	Extracellular Region	Yes

Majority of the un-annotated hypothetical genes in Chromosome 2 are localized in Nucleus and the remaining are localized in extracellular region, Mitochondria and cytoplasm. But it is interesting to note that one of the un-annotated hypothetical genes in Chromosome 2 (C2ORF66) contains transmembrane topology and localized in the extracellular region.

Conclusion

The present work has resulted in the identification of function for majority of the hypothetical proteins in the Human Chromosome 2 which can be validated experimentally Also, the functions predicted from the study gives a strong belief that the probability of expression of these hypothetical genes is very high, but further study is essential to know their expression condition. Another important aspect of such hypothetical ORFs also gives a clue that malfunctioning of similar functioning protein may enhance the expression of these

hypothetical genes and act as a standby gene for the functioning of the organism. The remaining un-annotated hypothetical genes were attempted to predict the topology and sub-cellular localization which gives preliminary clues to understand its function. On the other important techniques such as co-expression patterns analysis and phylogenetic profiling can be employed to understand the function of such genes.

References

[1.] Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talon and Montserrat Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics, 2005, Volume 21, Issue 18 Pp. 3674-3676.

[2.] Bhattacharya, A., Lakhman, S.S., Singh, S. (2004). Modulation of L-type calcium channels in Drosophila via a pituitary adenylyl cyclase-activating polypeptide (PACAP)-mediated pathway. J. Biol. Chem. 279(36): 37291--37297.

[3.] Chen, Y. and Xu, D. (2003) Computation analysis of high-throughput protein-protein interaction data. *Current Peptide and Protein Science*, 4, 159-181.

[4.] Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH. The NCBI BioSystems database. *Nucleic Acids Res.* 2010 Jan; 38(Database issue):D492-6. *(Epub 2009 Oct 23)*

[5.] Human epithelial cells trigger dendritic cell mediated allergic inflammation by producing TSLP.

[6.] Krogh A, Larsson B, von Heijne G, Sonnhammer EL Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001 Jan 19;305(3):567-80.

[7.] Lei Kong, Yong Zhang, Zhi-Qiang Ye, Xiao-Qiao Liu, Shu-Qi Zhao, Liping Wei and Ge Gao, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine, Nucleic Acids Research, 2007, Volume 35, Issue suppl 2 Pp. W345-W349

[8.] Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.

[9.] Nakai K and Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization, Trends Biochem Sci. 1999 Jan;24(1):34-6.

[10.] Nat Immunol. 2002 Jul;3(7):673-80. Epub 2002 Jun 10.

[11.] Pandit SB, Gosar D, Abhiman S, Sujatha S, Dixit SS, Mhatre NS, Sowdhamini R, Srinivasan N, SUPFAM--a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. Nucleic Acids Res. 2002 Jan 1;30(1):289-93.

[12.] Robert D. Finn, Alex Bateman, Jody Clements, Penelope Coggill ,Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, Erik L. L. Sonnhammer, John Tate and Marco Punta, Pfam: the protein families database, Nucleic Acids Research, 2013, Volume 42, Issue D1, Pp. D222-D230.

[13.] Roy, N. S., Farheen, S., Roy, N., Sengupta, S. and Majumder, P. P. (2008), Portability of Tag SNPs Across Isolated Population Groups: An Example from India. Annals of Human Genetics, 72: 82–89.

[14.] Shu-Ye Jiang1, Alan Christoffels2, Rengasamy Ramamoorthy1, and Srinivasan Ramachandran, Expansion Mechanisms and Functional Annotations of Hypothetical Genes in Rice Genome Plant Physiology Preview. Published on June 17, 2009, as DOI:10.1104/pp.109.139402

[15.] Soumelis V, Reche PA, Kanzler H, Yuan W, Edward G, Homey B, Gilliet M, Ho S, Antonenko S, Lauerma A, Smith K, Gorman D, Zurawski S, Abrams J, Menon S, McClanahan T, de Waal-Malefyt Rd R, Bazan F, Kastelein RA, Liu YJ

[16.] Tobias, J.A., Bates, J.M., Hackett, S.J. & Seddon, N. 2008. Comment on the latitudinal gradient in recent speciation and extinction rates of birds and mammals. Science 319: 901.

[17.] Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES and Kellis M, Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals, Nature. 2005 Mar 17;434(7031):338-45. Epub 2005 Feb 27.

[18.] Zarembinski, T.I., Hung, L.W., Mueller-Dieckmann, H.J., Kim, K.K., Yokota, H., Kim, R., and Kim, S.H. 1998. Structure-based assignment of the biochemical function of a hypothetical protein: A test case of Structural Genomics. Proc. Natl. Acad. Sci. 95: 15189–15193