Bioinformation
www.bioinformation.net

open access
Database

A database for human Y chromosome protein data

Pallipalayam Periyasamy Karthikeyan[1,*], Palaniswamy Thanga Velan Lakshmi[1], Chinmay Kumar Dwibedi[2], Arunachalam Annamalai[3]

[1]Phytomatics Laboratory, Bioinformatics Division, Bharathiar University, Coimbatore, India; [2]Bioinformatics division, School of Biosciences and Technology, VIT University, Vellore, India; [3]Department of Biotechnology, School of Biotechnology & Health Science, Karunya University, Coimbatore; Pallipalayam Periyasamy Karthikeyan – Email: ppkarthikeyan@gmail.com; * Corresponding Author

Received October 06, 2009; Revised October 16, 2009; Accepted October 20, 2009; Published October 24, 2009

Abstract:
The human Y chromosome is the sex determining chromosome. The number of proteins associated with this chromosome is 196 and 107 of the 196 proteins have yet not been characterised. Here, we describe the analysis of these 107 proteins by computing various physico-chemical properties using sequence and predicted structural data to elucidate molecular function. We present the derived data in the form a form a database made freely available for download, review, refinement and update.

Keywords: Y chromosome protein; human; homology modelling; sequence; function

Availability: http://puratham.googlepages.com/ [or] http://puratham.googlepages.com/ftpconnection

Background:
The presence of the Y chromosome determines the male characteristics in a mammalian embryo **[1]**. It is one of the smallest chromosomes in the human genome (~60Mb) with a limited number of genes **[2]**. The human Y chromosome comprises 59 million base pairs approximately (59,373,566 bps) with 61 known protein-coding genes, 25 novel protein-coding genes, 282 pseudogenes, 15 miRNA genes, 6 rRNA gens,13 snRNA genes, 1 snoRNA genes, 1 Misc RNA genes showing about 91,437 SNPs. The complete sequence of the male specific region of Y chromosome (MSY) comprising 95% of the chromosome length revealed about 78 protein-coding genes and about 27 are distinct proteins **[3]**. Y chromosome loss and rearrangements have been associated with different types of oncogenic disorders like male sex cord stroma tumor **[4]**, lung cancer **[5]**, esophageal carcinoma **[6]**, germ cell tumor **[7]**, turners syndrome **[8]** and bladder cancer **[9]**. Both oncogenes and tumour suppressor genes are hypothesised be present in this chromosome causing genetic disorders in male-specific organs such as testis **[2]**. The human male infertility has been attributed to mutations in the genes on Y chromosome **[10]**. Genetic or inherited disease or specific abnormalities in the Y chromosome are major factors for male infertility. Infertility men reveal many abnormal conditions, which include azoospermia, oligozoospermia, teratozoospermia, asthenozoospermia, necrospermia and pyospermia **[11]**. Despite its central role in sex determination, genetic analysis of the Y chromosome has been limited due to the paucity of available genetic markers **[12]**. MSY genes participate in diverse processes such as skeletal growth, germ cell tumorigenesis, graft rejection, gonadal sex determination, and spermatogenic failure **[13]**. A study on the function prediction of the 107 hypothetical proteins of the Y chromosome proteins is performed. Here, we describe the use of prediction methods to characterize the unknown functional information using sequence and modelled structural data and store derived data in the form of a database.

Methodology:
Dataset:
Homo sapiens Y chromosome protein sequences were retrieved from the EBI-EMBL database in FASTA format. A total of 196 proteins are available (May 2008) and 107 protein sequences have not yet been characterized.

Sequence analysis:
Sequence analysis was done using protein sequence analysis tools such as COMPUTE Pi/Mw **[14]**, PROTPARAM **[15]** & RADAR **[16]**. Various physical and chemical properties such as molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half life, instability index, aliphatic index and grand average of hydropathicity (GRAVY) were computed. Internal duplication and alignment of repeats were predicted using RADAR.

Secondary structure analysis:
Secondary structure prediction analysis was performed using online tools such as GOR IV (Garnier-Osguthorpe-Robson IV) **[17]**, HNN (Hierarchial Neural Network) **[18]** and SOPMA (Self Optimized Prediction Method with Alignment) **[19]**.

Template search:
Template is selected by BLAST search against protein databank (PDB) with > 40% sequence identity cut off **[20]**.

Tertiary structure analysis:
Tertiary structure prediction was performed using I-Tasser Server for *ab initio* structure prediction **[21]** and SWISS-MODEL for homology modelling **[22]**. The 3D structures modelled were visualized using PYMOL.

Structure validation:
Structure validations for the modelled 3D structures were done by using SAVS server (structure analysis and verification servers) for PROCHECK analysis **[23]**, Artificial Intelligence Decoys Evaluator (AIDE) **[24]** and ProQ **[25]**.

Function prediction:
The query protein sequences were subjected to BLAST **[20]**, INTERPROSCAN **[26]**, COG **[27]**, and PFAM **[28]** to assign predicted biochemical and cellular functions. The results were analyzed based on the confidence level of 25% using default parameters. Results from different tools were summed up to calculate 100% reliability.

Database:
We present the derived data in the form a form a database made freely available for download, review, refinement and update.

Database features:
The human Y chromosome has been studied for more than 30 years. It is used as a powerful tool to study human population and evolutionary data. The most characterizing feature of this chromosome is in human sex determination and in male germ cell development and maintenance. Therefore, it is important to document physical and chemical properties, sequence comparison, secondary structures, folding class, tertiary structure and biochemical information. Properties such as molecular weight, theoretical pI, estimated half life, extinction coefficients, instability index, aliphatic index, grand average of hydro-pathicity along with the number of negatively and positively charged residues were also document for each entry.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(5): 184-186 (2009)

Weight and atoms:
In the dataset, the entry A8MQV7 has the highest no of atoms 16175 with molecular weight 114705.1 and highest number of negatively charged residues (139). On the other hand, the entry Q6KER0 has the lowest total no of atoms and molecular weight of 159 and 1132.3, respectively.

Theoretical pI:
Q13381 was found to have the highest theoretical pI of 12.31, while A6NMP8 was found to have the lowest theoretical pI 3.29.

Half life:
Half life is the predicted time it takes for half of the amount of protein in a cell to disappear after synthesis. An acceptable value of 30 half life was found in all the proteins.

Extinction coefficient:
Extinction coefficient indicates the light absorbed by proteins at a certain wavelength. A6NDE4 in the dataset was found to have the highest extinction co-efficient value of 79900 while and A6NM12 showed the least value of 0.

Instability Index:
The instability index provided the stability of a protein. A highest value of 142.90 and lowest value of -51.21 was recorded for entries O14606 and A8MWL0, respectively in the dataset.

Aliphatic Index:
The aliphatic index of a protein was the relative volume occupied by the aliphatic side chains (alanine, valine, isoleucine and leucine). The entry A8MW33 has the highest aliphatic index of 136.53 and a lowest value of 7.22 for A6NMP8.

GRAVY:
The GRAVY value for a peptide or protein was calculated as the sum of hydropathy values of all the amino acids divided by the number of residues in the sequence. The maximum (0.860) and minimum (-1.404) values are recorder for entries A2RUG3 and A6NDE4, respectively in the dataset.

Charged residues:
The highest number of negatively charged amino acids (139) was found in A8MQV7 and the number of positively charged residues was found to be highest in Q24JR0 (94). A NIL value is recorded for the entry A6NMP8.

Predominant residues:
The frequencies of individual residues were calculated in terms of their percentages for each entry. The entry Q13381 had the highest percentage of amino acid serine (23.8%) followed by arginine with 22.2% in Q6KERO.

Atomic composition:
Atomic compositions each protein was calculated by PROTPARAM. The entry A8MQV7 has the highest composition of major atoms while and least number is recorder for Q6KER0.

Repeats:
Analysis shows that 39 protein sequences in the dataset showed the presence of repeats. This is not true with the remaining entries in the dataset. Among those with repeats, the entry Q24JR0 showed the highest number (20) of repeats at positions (9-718) with different scores.

Secondary structures:
Secondary structure prediction was done using GOR, HNN & SOPMA. The entry Q8N4A2 has the highest percentage of alpha helix in the dataset. Q6KER0, A6NNB5 and A6NII1 have the highest percentage of extended strand while Q9BZ97 has the lowest percentage.

Random coils:
A6NII1 has the highest percentage of random coil while A6NNB5 has the lowest percentage of random coil.

Homology models:
Homology modelling was done using SWISS MODELLER. The templates (A6NLN9 -2ZCY, A6NJH9-1D7Q, A6NMX1-1LL2, A8MTA1-2I13, A8MUG6-2NZ2, Q13369-2FY1, Q13374-2FY1) were used to model the 3D structures for 7 entries. The remaining 100 entries were modelled using the I-Tasser server.

Functional interpretation:
The sequences in the dataset were subjected to BLAST, INTERPROSCAN, COG, and PFAM to predict function. In this analysis, 10 proteins (A6NDE4, A6NEQ7, A6NJH9, A6NMU8, A8MUG6, Q8TD47, Q496E4, P22090, P0C7P1, and A6NEQ0) were identified to produce 100% reliable functions such as RNA recognition motif, ankyrin, eukaryotic initiation factor, neuroligin, argininosuccinate synthase, ribosomal protein S4E and RNA binding proteins, respectively. However, 24 different proteins were predicted with 75% confidence levels, among which, 11 proteins (A6NEC3, A6NFC0, A6NMT9, A8MZ49, A8MU69, A6NFK1, A6NGF5, A6NGT6, A6NDJ3, A6NJD2, A4FUW6) to have a unique function of nucleosome assemble protein, while 4 proteins (A6NMX1, A6NLN9, A6NHG5, A6NIC0) are of glycosyl transferase family 8 function. The remaining 9 proteins had different function with entry A6NCS7 for Tetratricopeptide repeat, Q24JR0 for Zfx / Zfy transcription activation region, O43610 for Sprouty protein, A8MUH9 for Y chromosome RNA recognition motif, A8MT17 for zinc finger, A8MQV7 for serine protease Inhibitor, A6NJK3 for tektin, A6NJY1 for sodium/hydrogen exchanger family and Q99218-2 for amelogenin.

Conclusion:
More than hundred human Y chromosome proteins (107) have not yet been characterized. We used I-Tasser to model 100 structures and SWISS MODELER for the remaining 7 proteins. 39 protein sequences indicate the internal repeats in RADAR and function prediction with 10 proteins (100% confidence), 24 proteins (75% confidence), 12 proteins (50% confidence) and 14 proteins (25% confidence). Analysis shows that most proteins are similar to ribosomal protein S4E, nucleosome assembly protein (NAP) and RNA-binding proteins. We present these data in the form a web database made available freely over the internet.

Acknowledgement:
I express my sincere thanks to Prof. Yang Zhang of Kansas University for his inspiration in this project and gratitude to Ambrish Roy (Kansas University), Dr. S. Parthasarathy (Bharathidasan University), K.B. Pavithra & S. Radhika (Bharathiar University) for their moral support.

References

[1] DC Page *et al., Cell* **51**:1091 (1987) [PMID: 3690661]
[2] L Quintana-Murci, M Fellous, *J Biomed Biotechnol.* **1**:18 (2001) [PMID: 12488622]
[3] K Ginalski *et al., Proc Natl Acad Sci U S A.* **101**: 2305 (2004) [PMID: 14983005]
[4] WF De Graaff, *et al., Cancer Genet Cytogenet.* **112**: 21 (1999) [PMID: 10432930]
[5] R Center *et al., Int J Cancer.* **55**: 390 (1993) [PMID: 8397161]
[6] S Hunter *et al., Genes Chromosomes Cancer* **8**:172 (1993) [PMID: 7509625]
[7] KL Nathanson *et al., Am J Hum Genet.* **77**:1034 (2005) [PMID: 16380914]
[8] C Ravel, JP Siffroi, *Gynecol Obstet Fertil.* **37**: 511 (2009) [PMID: 19464936]
[9] G Sauter *et al., Cancer Genet Cytogenet.* **82**: 163 (1995) [PMID: 7664248]
[10] S Ali, SE Hasnain, *Gene* **321**: 25 (2003) [PMID: 14636989]
[11] J Poongothai *et al., Singapore Med J.* **50**: 336 (2009) [PMID: 19421675]
[12] P Goodfellow *et al., J Med Genet.* **22**: 329 (1985) [PMID: 3908683]
[13] PH Vogt *et al., Cytogenet Cell Genet.* **79**:1 (1997) [PMID: 9533010]

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(5): 184-186 (2009)

[14] B Bjellqvist *et al., Electrophoresis* **14**:1023 (1993) [PMID: 8125050]
[15] MR Wilkins et al., *Methods Mol Biol.* **112**:531-52 (1999) [PMID: 10027275]
[16] A Heger, L Holm, *Proteins* **41**: 224 (2000) [PMID: 10966575]
[17] J Garnier *et al., Methods in Enzymology* **266**: 540
[18] Y Guermeur *et al., Bioinformatics* **15**: 413 (1999) [PMID: 10366661]
[19] C Geourjon *et al., Comput Appl Biosci.* **11**: 681 (1995) [PMID: 8808585]
[20] SF Altschul *et al., J Mol Biol.* **215**: 403 (1990) [PMID: 2231712]
[21] Y Zhang *BMC Bioinformatics* **9**: 40 (2008) [PMID: 18215316]
[22] K Arnold *et al., Bioinformatics* **22**: 195 (2006) [PMID: 16301204]
[23] RA Laskowski *et al., J Appl Cryst.* **26**: 283 (1993)
[24] P Mereghetti *et al., BMC Bioinformatics* **9**: 66 (2008) [PMID: 18230168]
[25] B Wallner, A Elofsson, *Protein Sci.* **12**:1073 (2003) [PMID: 12717029]
[26] E Quevillon *et al., Nucleic Acids Res.* **33**: W116 [PMID:15980438]
[27] RL Tatusov *et al., Nucleic Acids Res.* **29**: 22 [PMID: 11125040]
[28] RD Finn *et al., Nucleic Acids Res.* **36**: D281 (2007) [PMID: 18039703]

Edited by P. Kangueane
Citation: Karthikeyan *et al.*, Bioinformation 4(5): 184-186 (2009)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(5): 184-186 (2009)