Molecular Database Generation for Type 2 Diabetes using Computational Science-Bioinformatics' Tools

Gagandeep Kaur Grewal
Dept of Computer Engineering, UCOE,
Punjabi University,Patiala
Punjab,India
gdeepgrewal@gmail.com

Dr. Amardeep Singh#
Dept of Computer Engineering, UCOE,
Punjabi University,Patiala
Punjab,India

Abstract— In this paper a new algorithm GIGC is proposed which is the modified form of glucose insulin meal GIM model. Diabetes mellitus is one of the worst diseases that are affecting adversely large population. This motivates many researchers to study the glucose-insulin endocrine regulatory system and to design algorithms that could be applied on real time patient and benefit the medical field. Diabetes occurs when blood glucose levels are too higher than normal. Type two diabetes, also known as mellitus has exceeded its growth to a large extent worldwide, and its impact on global health care problem has increased the interest of the scientific community Variation in insulin sequence is combined to glucose insulin meal simulation model to develop a modified algorithm. Glucose insulin meal simulation software doesn't simulate for T2D. The new algorithm GIGC has been implemented using glucose insulin meal software in which genetic variation along with insulin glucose secretion dynamically are combined to help in reducing the prevalence of T2D.

Keywords- BLAST, GIGC, NCBI, SWISS PROT, T2D.

I. INTRODUCTION

Major research efforts in Bioinformatics include sequence alignment, gene finding, drug design, drugs discovery, protein structure alignment, prediction of gene expression etc.[3].Since Diabetes mellitus is one of the worst diseases that are affecting adversely large population this motivates many researchers to study the glucose-insulin endocrine regulatory system. Life is difficult to diabetic patients. They must measure their glucose rate, inject insulin regularly, visit physician and examine results which are difficult to understand. Proposed simulation model GIGC may help them to minimize measurement time for detecting glucose level. By this way, insulin dosage may be planned effectively. An estimation that 50.8 million people are living with diabetes, India has the world's largest diabetes population, followed by China with 43.2 million (Source: IDF, Diabetes Atlas, 4th edition Last updated 4-27-2011 by bisl.wdf).But unfortunately most of diabetic patients either do not visit physician regularly or do not know he has diabetes already.

Our starting point to develop algorithm is to help these patients to understand their diabetic graphs, and also help physicians during medical treatment for dosage planning and this information is combined with genetic information i.e. about varying insulin that predicts where insulin is different from normal patient. Attention to such factors as insulin, glucose and meal, combined with genetic data identifying genetic risk factors, will help in reducing the prevalence of T2D (Type 2 diabetes) and can help medical field. For achieving this objective a case study of diabetes type 2 is taken. NCBI's database is used. Later BLAST searching algorithm is applied to find the sequence match between human insulin and its variants.

II. LITERATURE SURVEY

Some related works about diabetes and simulation models are found in literature survey and mentioned in this section. Studies reveal that the mostly used and also the simplest one is the minimal model of Bergman (bergman et al., 1979) for type 1 diabetes patients under intensive care, and its extension ,the three-state minimal model(bergman et al.,1981). But its simplicity proved to be its disadvantage too, as it is very sensitive to parameters variance, the plasma insulin concentration must be known as a function of time and in its formulation a lot of components of the glucose insulin interaction were neglected. But two models were helpful in developing new algorithm. in 2007[5], meal simulation model for the glucose insulin system in body was developed by Man, C., Rizza, R., &Cobelli,C. Model results describe both a single meal and daily life meal (three times a day) breakfast, lunch, dinner in normal.

In 2006, mixed meal simulation model of glucose insulin system was developed by Man, C., Rizza, R., & Cobelli, C.[6] Model results show normal daily life (breakfast, lunch, dinner) both in normal and pathophysiological situations. It simulates both open- and closed-loop insulin infusion strategies.

Several software exists, some are freely available, and some are commercial.Few of these are mentioned in the review paper by Lehmann and Deutsch (1995).Here is a partial list: (The URLs were last accessed on 27 April 2011).

AIDA*-http://www.2aida.org/aida/site-map.htm,*

WINSTODEC,

FSIGTT- *ftp://ftp.mbs.ele.tue.nl/CS/Riel/,*

DIAS*-*(*http://www.mi.auc.dk/~okh/dias.htm*)**,**

Glucosim (*http://216.47.139.198/glucosim/index.html*)

III. DIABETES

Diabetes is a chronic metabolic disorder that adversely affects the body's ability to manufacture and use insulin, a hormone necessary for the conversion of food into energy. The disease greatly increases the risk of blindness, heart disease, kidney failure, neurological disease.Type 2 diabetes (T2D) is a disease in which insulin is abnormally secreted or does not act correctly, leading to elevated blood glucose [1]. Over time, elevated glucose levels can lead to multiple organ damage. Diabetes is the leading cause of chronic renal failure, adult blindness, and limb amputation, and is a major risk factor for heart disease, stroke, and birth defects [2].T2D is believed to be a multi-factorial disease, i.e., it is influenced by both genetic and environmental factors. People with a family history of the disease are at higher risk of developing it themselves since they share genetic background and likely share similar environments. It has been estimated that 30%-70% of T2D risk can be attributed to genetics, with multiple genes involved and different combinations of genes playing roles in different subsets of individuals [2]. It is not yet known how many genes are involved or how much control each exerts over the development of the disease, but recent research has identified a number of promising candidates [2].

IV. BIOINFORMATICS IS APPLIED TO ATLEAST FIVE MAJOR TYPES OF ACTIVITIES[7]

1) Data acquisition: Data(related to the DNA samples, such as the species, tissue type, and quality parameters used in the experiments) acquisition is primarily concerned with accessing and storing data generated directly off of laboratory instruments.
2) Database development: Many laboratories generate large volumes of such data as DNA sequences, gene expression information, three-dimensional molecular structure, and high-throughput screening. Hence, they must develop effective databases for storing and quickly accessing data.
3) Data analysis: Bioinformatics analysts may write specific algorithms to analyze data, or they may be expert users of analysis tools, helping scientists understand how the tools analyze the data and how to interpret results.
4) Data Integration: Once information has been analyzed, a researcher often needs to associate or integrate it with related data from other databases.
5)Analysis of integrated data: A good deal of the early work in bioinformatics focused on processing and analyzing gene and protein sequences catalogued in databases such as GenBank, EMBL, and SWISS-PROT.

V. IMPLEMENTING DETAILS FOR THE PROPOSED MODEL

GIGC (Genetically insulin glucose control) is implemented using MATLAB 7.0. Insulin is a protein so sequence of amino acids of insulin is studied. Data about variants in human insulin and its variation in diabetic patient was generated. This is combined with the glucose insulin dynamics. This new algorithm “Genetically Insulin Glucose Control (GIGC)” has large possibilities of identifying candidate gene’s percentage variation in normal subject to type 2 diabetic patient. This model implements a variation of a model described by Dalla Man,Cobelli C.[8] In GIGC , algorithm is modified such that gene sequence for insulin is used to match the query (i.e. variant insulin) with actual sequence. The proinsulin precursor of insulin is encoded by INS gene.

Step sequence in GIGC is the following:

1. The user selects type 1, 2 or normal subject and sets the values for all fields asked, basal value is calculated. Simulation starts and profile is saved.

2. Now suppose user wants to compare type two values with normal, he simply has to click type two and click simulation button, then he is asked for which profile should be compared with; then user can enter type 1 or normal. Hence a graphical window is displayed with current and previous values.

3. MATLAB opens web browser in another window which displays the investigated protein sequence for insulin using NCBI BLAST algorithm in FASTA format. Since protein insulin is important in regulation of blood sugar and in diabetes, so its genetic variation must be kept in consideration to find new genetic treatments for type 2 diabetes. Swiss-Prot Protein Knowledgebase is visited. Enter ‘insulin precursor' into the Search box at the top of the page.

4 .On the results page, scroll down and click on INS HUMAN.

5. Under comments section, purpose of this protein is mentioned. A defect in Insulin causes *familial hyperproinsulinemia*. It shows Defects in INSR may be associated with noninsulin-dependent diabetes mellitus (NIDDM); also known as diabetes mellitus type 2.

6. Scroll down to the Features section. Locate the sections that contain variants. Some of these cause diseases such as diabetes. Click on one of the variants.

7. Return to the Human Insulin page. Scroll all the way to the bottom and look at Sequence Information. How long is the protein molecule? Click on the FASTA format link to get the sequence in a format we can use.

8. Highlight and copy the amino acid sequence:

```
>sp|P06213|INSR_HUMAN Insulin receptor OS=Homo sapiens GN=INSR PE=1 SV=4
MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHL
```

9. Go to the National Center for Biotechnology Information (NCBI) home. This database contains virtually all sequenced genes. Compare the human insulin protein sequence to everything previously sequenced. At the top of the page, click on BLAST (Basic Local Alignment Search Tool) and then on the next page “Protein-protein (blastp)”.

10. Click Format to see if the results are ready. After clicking once, the results page will refresh itself until the search is complete, this will take a few minutes.

11. The Colour Key for Alignment Scores shows how well the sequences found by BLAST match the sequence that one has entered.

12. Scroll down to see a list of the proteins with their E-values. E-Values are a mathematical representation of how related other proteins are to your search sequence. The smaller the number, the closer the match.

13. Below this list are the actual sequences. The top row is the insulin sequence searched for. The bottom row contains the matches found in the database. The middle row shows the parts that match. The percent that match is also given.

The annotated protein sequence for human insulin is studied, variants in sequence is retrieved using FASTA. Then at NCBI, BLASTp is performed to match the sequence for our query with previously sequenced insulin protein. This shows how much %age of diabetic individual’s insulin amino acid length is varied from normal patient. MATLAB programming is used for implementation of proposed algorithm. GIGC (Genetically Insulin Glucose Control) has been designed and implemented for type 2 diabetes using GIM simulation software [4]. This algorithm helps to plan insulin dosage effectively and also helps to know variant in insulin. For insulin’s genetics; NCBI database and BLAST were used. Hence, GIGC generates molecular database which can help diabetic patients.[9].

VI. SIMULATION RESULTS AND DISCUSSION

Glucose and insulin concentrations, glucose production, glucose utilization, meal rate of appearance, and insulin secretion rate obtained. Red line shows current solution and blue line shows previous solution.

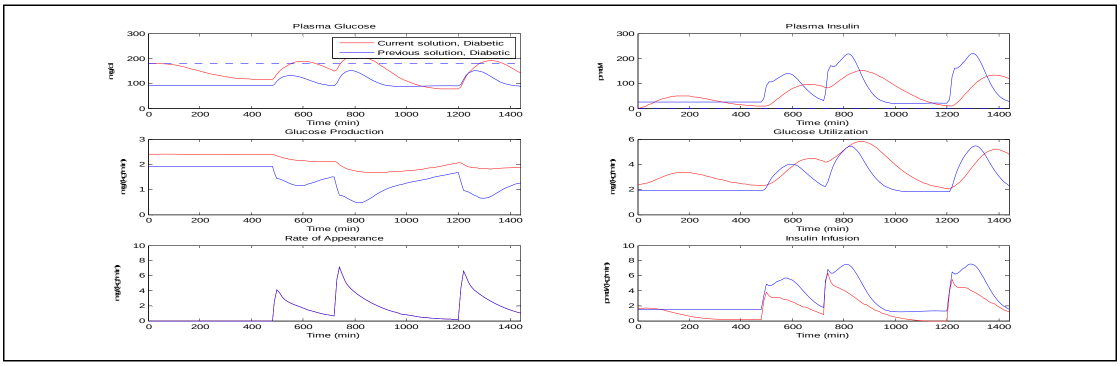


Figure 1. Represents simulation result for comparison of type 2 diabetes with normal.

The graphs below show the simulation of subject type 1 diabetes with subject type 2 diabetes. It compares the glucose production, glucose utilization, insulin infusion wrt. time.

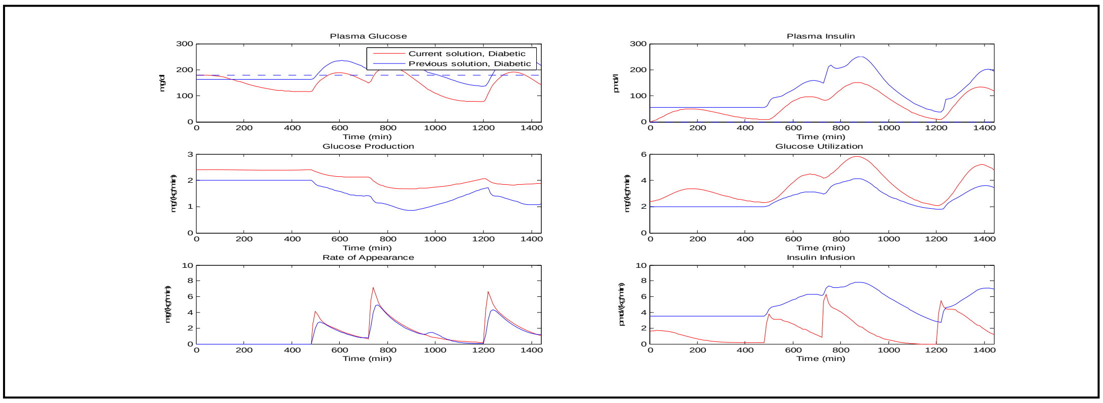
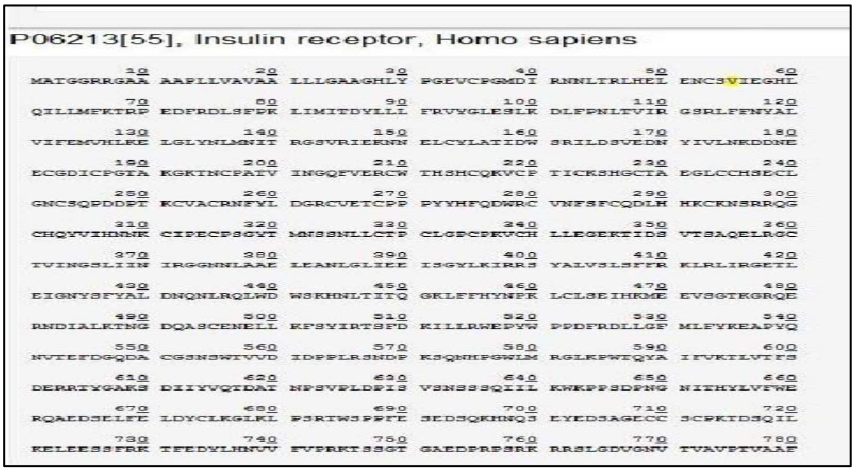


Figure 2. Simulation results compare diabetes type 1 with diabetes type 2 here.

Below screen shot shows that some of these variations in human insulin cause diseases such as diabetes.

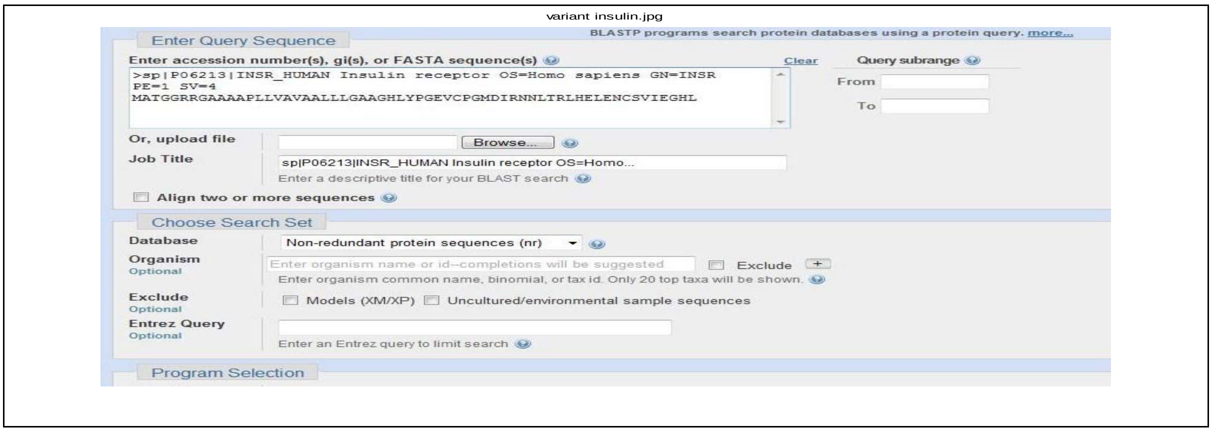
variant insulin.jpg


P06213[55], Insulin receptor, Homo sapiens

```
        10         20         30         40         50         60
MATGGRRGAA AAPLLVAVAA LLLGAAGHLY PGEVCPGMDI RNNLTRLHEL ENCSVIEGHL
        70         80         90        100        110        120
QILLMFKTRP EDFRDLSFPK LIMITDYLLL FRVYGLESLK DLFPNLTVIR GSRLFFNYAL
       130        140        150        160        170        180
VIFEMVHLKE LGLYNLMNIT RGSVRIEKNN ELCYLATIDW SRILDSVEDN YIVLNKDDNE
       190        200        210        220        230        240
ECGDICPGTA KGKTNCPATV INGQFVERCW THSHCQKVCP TICKSHGCTA EGLCCHSECL
       250        260        270        280        290        300
GNCSQPDDPT KCVACRNFYL DGRCVETCPP PYYHFQDWRC VNFSFCQDLH HKCKNSRRQG
       310        320        330        340        350        360
CHQYVIHNNK CIPECPSGYT MNSSNLLCTP CLGPCPKVCH LLEGEKTIDS VTSAQELRGC
       370        380        390        400        410        420
TVINGSLIIN IRGGNNLAAE LEANLGLIEE ISGYLKIRRS YALVSLSFFR KLRLIRGETL
       430        440        450        460        470        480
EIGNYSFYAL DNQNLRQLWD WSKHNLTITQ GKLFFHYNPK LCLSEIHKME EVSGTKGRQE
       490        500        510        520        530        540
RNDIALKTNG DQASCENELL KFSYIRTSFD KILLRWEPYW PPDFRDLLGF MLFYKEAPYQ
       550        560        570        580        590        600
NVTEFDGQDA CGSNSWTVVD IDPPLRSNDP KSQNHPGWLM RGLKPWTQYA IFVKTLVTFS
       610        620        630        640        650        660
DERRTYGAKS DIIYVQTDAT NPSVPLDPIS VSNSSSQIIL KWKPPSDPNG NITHYLVFWE
       670        680        690        700        710        720
RQAEDSELFE LDYCLKGLKL PSRTWSPPFE SEDSQKHNQS EYEDSAGECC SCPKTDSQIL
       730        740        750        760        770        780
KELEESSFRK TFEDYLHNVV FVPRKTSSGT GAEDPRPSRK RRSLGDVGNV TVAVPTVAAF
```

Fig 3. Variants in human Insulin.

Go to the **National Center for Biotechnology Information** (NCBI) home page. This database contains virtually all sequenced genes. We will compare the human insulin protein sequence to everything previously sequenced.At the top of the page, click on "BLAST" and then on the next page "Protein-protein(blastp)".



variant insulin.jpg

Enter Query Sequence

BLASTP programs search protein databases using a protein query. more...

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>sp|P06213|INSR_HUMAN Insulin receptor OS=Homo sapiens GN=INSR
PE=1 SV=4
MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHL
```

Clear

Query subrange

From

To

Or, upload file Browse...

Job Title sp|P06213|INSR_HUMAN Insulin receptor OS=Homo...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

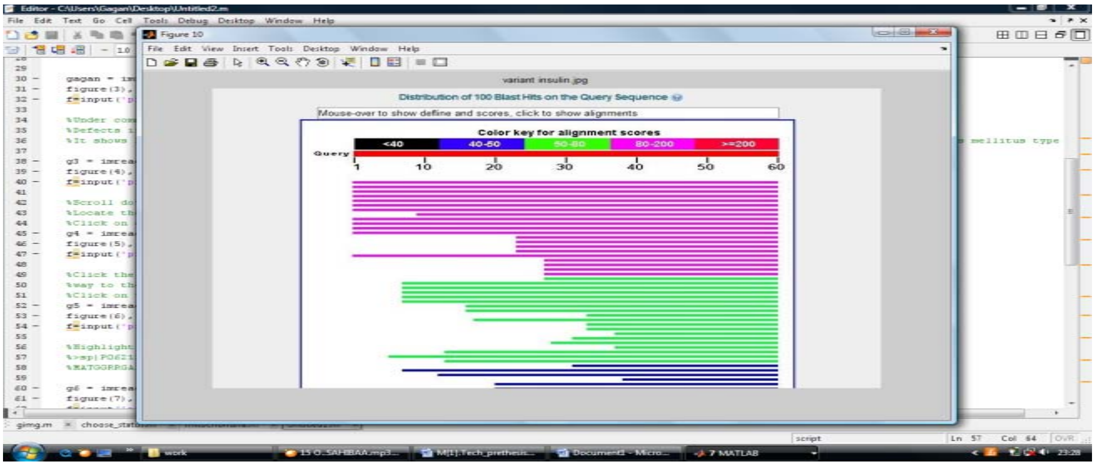
Entrez Query Optional:

Enter an Entrez query to limit search

Program Selection

Fig 4. BLASTp performed

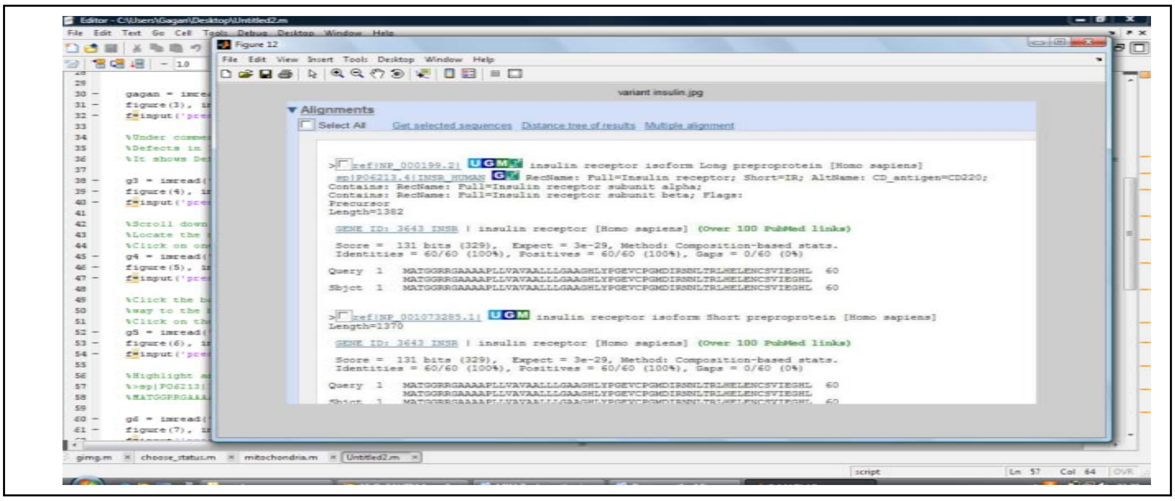
Below given figure shows how well the sequences found by BLAST match the sequence you entered. This is known as alignment score.



Editor - C:\Users\Gagan\Desktop\Untitled2.m
File Edit Text Go Cell Tools Debug Desktop Window Help
Figure 10
File Edit View Insert Tools Desktop Window Help
variant insulin.jpg
Distribution of 100 Blast Hits on the Query Sequence
Mouse-over to show define and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
10
20
30
40
50
60
gimg.m
choose_stat
script
Ln 57
Col 64
OVR
work
15 O_SAHIBAA.mp3...
M[1].Tech_prethesis...
Document1 - Micro...
7 MATLAB
23:28

Fig 5. The Color Key for Alignment Scores

In figure 6, the top row is the insulin sequence you searched for. The bottom row contains the matches found in the database. The middle row shows the parts that match. The percent that match is also given.



```
>ref|NP_000199.2| UGM insulin receptor isoform Long preproprotein [Homo sapiens]
 sp|P06213.4|INSR_HUMAN G RecName: Full=Insulin receptor; Short=IR; AltName: CD_antigen=CD220;
Contains: RecName: Full=Insulin receptor subunit alpha;
Contains: RecName: Full=Insulin receptor subunit beta; Flags:
Precursor
Length=1382

 GENE ID: 3643 INSR | insulin receptor [Homo sapiens] (Over 100 PubMed links)

 Score =  131 bits (329),  Expect = 3e-29, Method: Composition-based stats.
 Identities = 60/60 (100%), Positives = 60/60 (100%), Gaps = 0/60 (0%)

Query  1   MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHL  60
            MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHL
Sbjct  1   MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHL  60

>ref|NP_001073285.1| UGM insulin receptor isoform Short preproprotein [Homo sapiens]
Length=1370

 GENE ID: 3643 INSR | insulin receptor [Homo sapiens] (Over 100 PubMed links)

 Score =  131 bits (329),  Expect = 3e-29, Method: Composition-based stats.
 Identities = 60/60 (100%), Positives = 60/60 (100%), Gaps = 0/60 (0%)

Query  1   MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHL  60
            MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHL
Sbjct  1   MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHL  60
```

Fig6. Result displaying the actual sequences.

Simulation models of glucose –insulin control system have proven useful for tackling various aspects of physiology and of diabetic control. This combined algorithm based on meals collaborated with identification of amino acid variation in a person’s insulin will have a profound impact on the future of biomedical field. This can not only help in controlling diet and glucose insulin mutations, but also investigate the sequence of human insulin in every diabetic person.

VII. CONCLUSION

The main purpose of creating and maintaining such databases in commercial organizations is their importance in the process of drug discovery. Human insulin’s protein sequence is studied. It is found that defects in INSR may be associated with T2D. Section that contains variants is searched for. Sequence information is retrieved by using FASTA format. Then NCBI’s database which contains virtually all sequenced genes compared our query with previously sequenced human insulin sequences. BLASTp is performed to match the sequence we entered. Result shows an alignment score of 60% that means our query matched 60%. Hence, variations in *TCF7L2* are associated with impaired insulin secretion and increased hepatic glucose production, which may partially explain the development of T2D in people carrying *TCF7L2* variations.

Difficulties in identifying specific gene mutations that cause type 2 diabetes are due to number of genes that are involved in controlling our fuel intake and regulation, lifestyle of a person and environment in which he is living add up to risks of this disease, inherited lifestyle, poor eating habits and lack of exercise. If genetic testing becomes easily approachable then insulin genetics study combined with meal glucose insulin models can be a boon to biomedical field. GIGC algorithm based on meals collaborated with identification of amino acid variation in a person‘s insulin can improve diagnosis of type two diabetes. The Genetics of Type 2 Diabetes Mellitus will identify candidate genes that predispose people to develop Type-2 diabetes.

ACKNOWLEDGMENT

Gagandeep Kaur Grewal thanks Guide Dr. Amardeep Singh, for his guidance, Professor.Claudio Cobelli for providing the simulation software.

REFERENCES

[1] Source: genetics and functional genomics of type 2 diabetes mellitus by Ayo Toye & Dominique Gauguier; Genome Biol.2003;4(12):241.

[2] *“FACT SHEET - Type 2 Diabetes”* by national institutes of health (updated October 2010) (The National Institute of Diabetes and Digestive and Kidney Diseases http://www.niddk.nih.gov/)

[3] Grewal Gagandeep Kaur,Dr.Singh Amardeep, *Bioinformatics and its applications in biomedical and agriculture,ETCSIT-2011*,106-107.

[4] Dalla Man C,Cobelli C, Raimondo DM, Rizza RA,(2007) GIM, simulation software of meal glucose- insulin model. JDST Vol 1,Issue 3.

[5] Man, C., Rizza, R., & Cobelli, C. (2007). Meal Simulation Model of the Glucose-Insulin System. *IEEE Transactions on Biomedical Engineering*, *54*(10), 1740-1749. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17946394

[6] Dalla Man, C., Rizza, R. , & Cobelli, C. . (2006). Mixed meal simulation model of glucose-insulin system. *Conference Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society*, *1*(2), 307-10. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17946394

[7] Recupero, Anthony J. "*Bioinformatics*." Genetics. 2003. Retrieved June 19, 2011 from Encyclopedia.com:http://www.encyclopedia.com/doc/1G2-3406500028.html

[8] Dalla Man C,Cobelli C, Raimondo DM, Rizza RA,(2007) GIM, simulation software of meal glucose- insulin model. JDST Vol 1,Issue 3.

[9] Grewal G.K.,Dr.Amardeep S., “Generating Molecular Database Using Biocomputing Approach”, Research Cell: An International Journal of Engineering Sciences, In press 2011.

Periodicals and web resources:

[1] BMC Bioinformatics. BioMed Central. (2010-2011) www.biomedcentral.com/bmcbioinformatics

[2] In Silico Biology. Bioinformation Systems. 1998-2008.www.bioinfo.de/isb

[3] PLoS Computational Biology. International Society for Computational Biology. 2005-2008.http://compbiol.plosjournals.org/perlserv/?request=index-html&issn=1553-7358

[4] http://www.ncbi.nlm.nih.gov/genbank/

[5] http://www.clcbio.com/index.php?id=1046

[6] www.google.com

[7] Bioinformatics Organization.http://bioinformatics.org [accessed 2 April 2011]

A bioinformatics society open to everybody. Strong emphasis on open access to biological information as well as free and open source software. “Bioinformatics Web-Tools Collection”.

AUTHORS PROFILE

Grewal Gagandeep Kaur submitted her dissertation for Masters of Technology in CE. Research areas include Bioinformaitcs, Computational Sciences, Biomedical field. Published 7 papers in International and 1 in Natioanl conference.