Generating Molecular Database using BioComputing Approach

[1]Gagandeep Kaur Grewal, [2]Dr.Amardeep Singh,
Department of CE, UCoE, Punjabi University, Patiala
[1]gdeepgrewal@gmail.com

Abstract: *In this paper, case study of diabetes is taken and a new algorithm GIGC is proposed which is the modified form glucose insulin meal GIM model. 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 to design algorithms that could be applied on real time patient and benefit the medical field.In this paper, 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.*

1. Introduction

Bioinformatics is the field of Science in which Biology, Computer Science and Information Technology merge to form a single discipline. Major research efforts in this field include sequence alignment, gene finding, drug design, drugs discovery, protein structure alignment, prediction of gene expression etc. [3]. Our approach is to generate a Biological Database having the informalities about the Specific Gene \Protein that play a central role in the nature. 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.

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. This simulation model may help them to minimize measurement time for detecting glucose level. By this way, insulin dosage may be planned effectively. Medical literature and world diabetes foundation declare that 70% of the current cases of diabetes occur in low- and middle income countries. 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 and can help medical field.

2. Literature Survey

Any researcher would find algorithms and models about glucose insulin dynamics in diabetes on internet or scientific library. Some related works about diabetes and simulation models are found in literature survey and mentioned in this section. 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.

3. 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, and other conditions for the approximately 16 million Americans who are affected by it.

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].

3.1 Type 2 diabetes risk factors [7]:

- Being overweight
- Lack of exercise
- Having a family history of diabetes
- Giving birth to a baby weighing more than 9 pounds or being diagnosed with gestational diabetes
- Having high blood pressure
- Having an HDL, or "good," cholesterol level below 35 mg/dL
- Having a triglyceride level above 250 mg/dL
- Having polycystic ovarian syndrome
- Having heart disease
- A number of genes that influence susceptibility to type 2 diabetes have been identified, but no one gene is strongly associated.

4. Bioinformatics is applied to at least five major types of activities

1) **Data acquisition:** Data acquisition is primarily concerned with accessing and storing data generated directly off of laboratory instruments. The data had to be captured in the appropriate format, and it had to be capable of being linked to all the information related to the DNA samples, such as the species, tissue type, and quality parameters used in the experiments.

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. Consequently, they must develop effective databases for storing and quickly accessing data.

3) **Data analysis:** Bioinformatics analysts have a broad range of opportunities. They 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. For example, a scientist may run a series of gene expression analysis experiments and observe that a particular set of 100 genes is more highly expressed in cancerous lung tissue than in normal lung tissue. The scientist might wonder which of the genes is most likely to be truly related to the disease. To answer the question, the researcher might try to find out more information about those 100 genes, including any associated gene sequence, protein, enzyme, disease, metabolic pathways, or signal transduction pathway data. Such information will help the researcher narrow the list down to a smaller set of genes. Finding this information, however, requires connections or links between the different databases and a good way to present and store the information. An understanding of database architectures and the relationship between the various biological concepts in the databases is key to doing effective data integration.

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. Such databases were developed in academia or by government-sponsored groups and served as repositories where scientists could store and share their sequence data with other researchers. With the start of the Human Genome Project in 1990, efforts in bioinformatics intensified, rising to the challenge of handling the large amounts of DNA sequence data being generated at an unprecedented rate.

5. Implementing details

GIGC is implemented using MATLAB 7.0.

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. Side by side another window also opens 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.Defects in Insulin cause *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

   MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRN

   NLTRLHELENCSVIEGHL
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" 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.

6. 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.

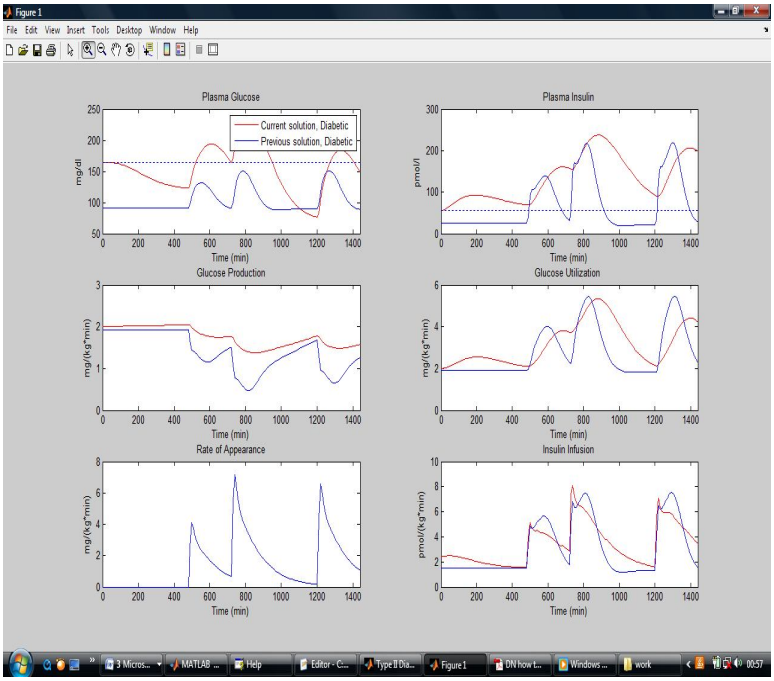

Figure 1. *Represents simulation results for comparison of type2 diabetes with normal.*

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

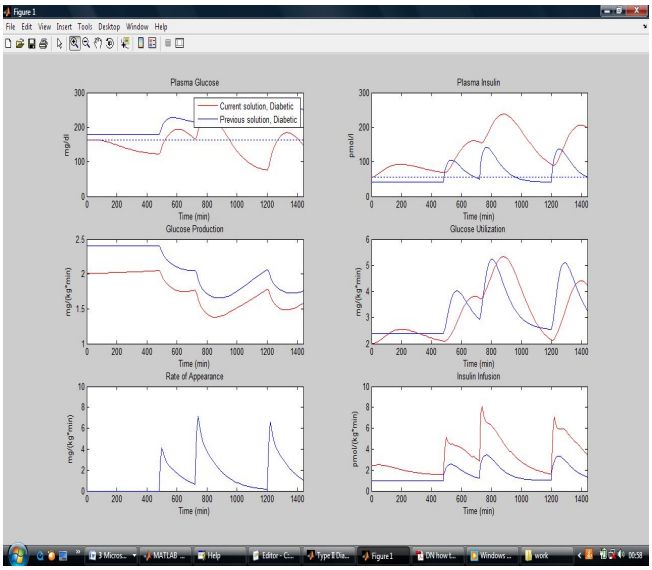

Figure 2. *Simulation results compare diabetes type 1 with diabetes type 2 here.*

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

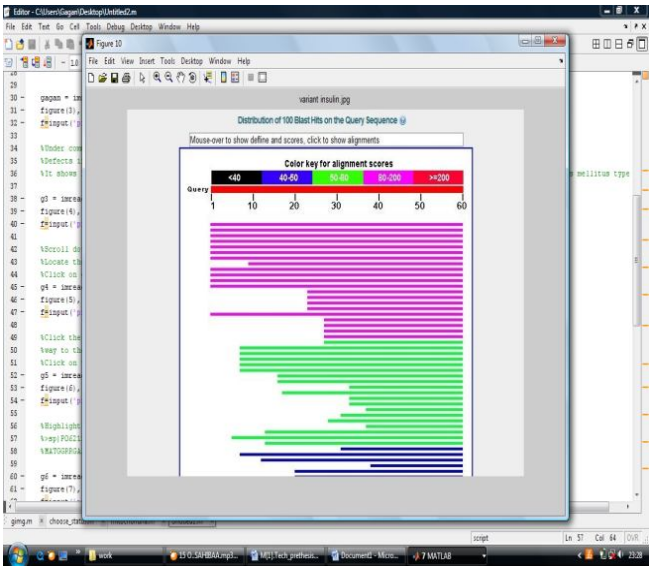

Figure 3 : *The Color Key for Alignment Scores*

In figure 4, 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.

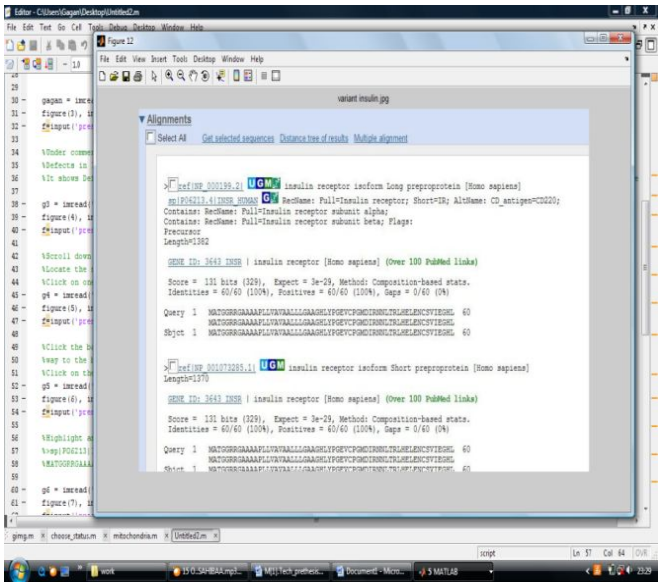

Figure 4 : *Result displaying the actual sequences.*

Simulation models of glucose –insulin control system have proven useful for tackling various aspects of physiology and of diabetic control. Since insulin resistance is frequently associated as predictor or pathogenic factor in different pathological conditions (type 2 diabetes mellitus, obesity, etc), it is important to have simple, accessible tools for evaluating insulin sensitivity of target tissues to the insulin action in humans. Since Type 2 diabetes mellitus arises from a deteriorated tissue response to biological effects of insulin and impaired glucose induced insulin secretion [1] 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. Following table shows the comparison between previously model and proposed model GIGC.

Table 1. Comparison of glucose insulin meal algorithm with genetically insulin glucose control algorithm

Glucose meal insulin model GIM	Genetically insulin Glucose control model GIGC
Complex	Easy to use and understand
No association with genetics of diabetes	Association with genes/proteins that increase the rate of diabetes
Impaired insulin secretion and increased hepatic glucose production results in peak rate for glucose production, utilization and insulin infusion.	Impaired insulin secretion and increased hepatic glucose production results in variations in TCF7L2
Only dynamics of meal, glucose, insulin is considered	Includes genes/protein variation (insulin in this case) of normal to diabetic patient
Doesn't simulate for T2D	Simulates for comparing T2D with normal and T1D and insulin sequence match.

7. Conclusion

Human insulin's protein sequence is studied. It is found that defects in INSR may be associated with T2D. Then the 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. We get 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.

If genetic testing becomes easily approachable then insulin genetics study combined with meal glucose insulin models can be a boon to biomedical field. This glucose insulin 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. Identification of genes and proteins that prevent the pancreas from producing proper amounts of insulin as well as those that prevent insulin from working properly in other tissues in the body.

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

[5] Diabetes Sources: MedlinePlus – *"Diabetes"*,Last Updated (Tuesday, 24 August 2010 18:27)

Periodicals:

[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

Web Resources:

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

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

[3] www.google.com

[4] 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".