Covenant Journal of Physical & Life Sciences (CJPL) Vol. 5 No. 2, Dec 2017

An Open Access Journal available online

Anopheles gambiae PGDB, AnoCyc, version 1.1 in Summary

Adebiyi M.[1*], Fatumo, S.[1] & Adebiyi E.[1]

[1]Covenant University Bioinformatics Research Cluster,
Department of Computer and Information Sciences, Covenant University, Ota,
*Correspondence: Adebiyi Marion, marion.adebiyi@covenantuniversity.edu.ng

***Abstract*:** Identification of novel insecticidal targets for the development of more effective insecticides is critical, discovery of new resistance mechanisms, computational analysis of biochemical pathways with various genomic data is essential for novel discoveries in proffering solution to the challenge of malaria vector eradication efforts world research community. Dissection and comprehensive study of biochemical metabolic networks has great potential to effectively and specifically identify essential enzymes as potential insecticidal targets. Using the PathoLogic program, we have constructed AnoCyc, version 1.1, a pathway/genome database (PGDB) for *A. gambiae* AgamP3, using its annotated genomic sequence and other annotated information from Vector base, UNIPROT and KEGG databases. AnoCyc for AgamP3, the first PGDB for A gambiae has been deployed to BioCyc Database collection, which has a total size of about 273,093,681pb with 1237 compounds, 3684 Go terms, 255 pathways, 2380 enzymes without any protein complex, 1758 enzymatic reactions, 42 super pathways and several other components and their corresponding information.

Introduction

The completion of the genome sequences for some organism including that of the deadliest malaria vector, *A. gambiae* [1] has given opportunity to develop various methods to facilitate effective malaria control strategies. Several works had attempted to use genomics to explain the mode of mosquito resistance, reveal essential reactions, genes and pathways and predict drug target [2].

Using the PathoLogic program [3], we constructed AnoCyc 1.0, a pathway/genome database (PGDB) for *A. gambiae* AgamP3, using its annotated genomic sequence and other annotated information from UNIPROT and KEGG databases. This formulated the very first edition and version of PGDB for the malaria vector, Mosquitoes. BioCyc

version 17.5 hosted this in December 2012.

BioCyc [4] is a collection of over 200 pathways/genome databases, containing whole databases dedicated to certain organisms. For instance, AnoCyc, which falls under the giant umbrella of BioCyc, is a highly detailed bioinformatics database on the genome and metabolic reconstruction of the *A. gambiae*. The AnoCyc database can serve as a model for any reconstruction, additionally; it is an encyclopaedia of metabolic pathways contains a wealth of information on metabolic reactions derived from over 600 different organisms including various species of *Anopheles, Plasmodium and Homo sapiens* [5]. BioCyc database has three main Tiers; this quality of a database detrmines the tier it belongs to. The most accurate of all databases are the Tier 1 databases, they are manually curated and according to [6], and they have received person-decades of literature based curation. Tiers 2 (36 PGDBs) and 3 (3521 PGDBs) databases are curated with the pathologic program; they contain computationally predicted metabolic pathways.

Review of Existing Literature

There is lots of information on various databases about metabolic pathways, reactions, the enzymes that catalyzes the reaction. A few of these databases and resources that are crucial and serve as backbones to metabolic construction and reconstruction were discussed.

Gene bank

The GenBank sequence database is an annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced at National Center for Biotechnology Information (NCBI) as part of an international collaboration with the European Molecular Biology Laboratory (EMBL) Data Library from the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ). GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. GenBank continues to grow at an exponential rate, doubling every 10 months. Release 134, produced in February 2003, and contained over 29.3 billion nucleotide bases in more than 23.0 million sequences. GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers [7].

AnoBase

AnoBase is a database containing genomic/biological information on *anopheline* mosquitoes, with an emphasis on *Anopheles gambiae*, the world's most important malaria vector. AnoBase is the continuation of AnoDB, a database established and maintained since 1996 at the Institute of Molecular Biology and Biotechnology (IMBB) of the Foundation of Research and Technology - Hellas (FORTH) in Heraklion, Crete, Greece. AnoBase was funded by grants from the UNDP/World Bank/World Health Organization Special Programme for Research and Training in Tropical Diseases (TDR) and is now supported by the National Institute of Allergy and Infectious Diseases (NIAID), as a member of VectorBase, that is, the database of insect disease vectors.

Kyoto encyclopedia of genes and genomes (KEGG)

KEGG is a bioinformatics database resource containing information on genes, proteins, reactions and pathways, for understanding high-level functions

and utilities of the biological system, such as the cell, the organism and the ecosystem. This resource is extremely useful when building association between metabolism, enzymes, reactions and genes. The KEGG database [8] consists of three databases that are tightly linked: PATHWAY given a network of interacting molecules, GENES hosts a collection of sequenced genomes and LIGANDS contains compounds, enzymes and enzymatic reactions.

Materials and Methods

Using the annotated genomic data of *A. gambiae* AgamP3 from the NCBI, and more annotations from UNIPROT and KEGG, metabolic network data of *A.gambiae* AgamP3 was built and lunched on BioCyc as AnoCyc 1.0 database, (http://www.biocyc.org, Version 17.1) [16]. The first version of AnoCyc, the PGDB for *A. gambiae* AgamP3 deployed under the www.bioCyc.org databases (http://biocyc.org/ANO2/organism-summary?object=ANO2) [9]. This version describes 14974 genes, of which 14324 code for polypeptides and 650 codes for RNA. There are 2297 known enzymes. Information on transport processes between the different compartments includes 115 transporter metabolites and 10 transport reactions.

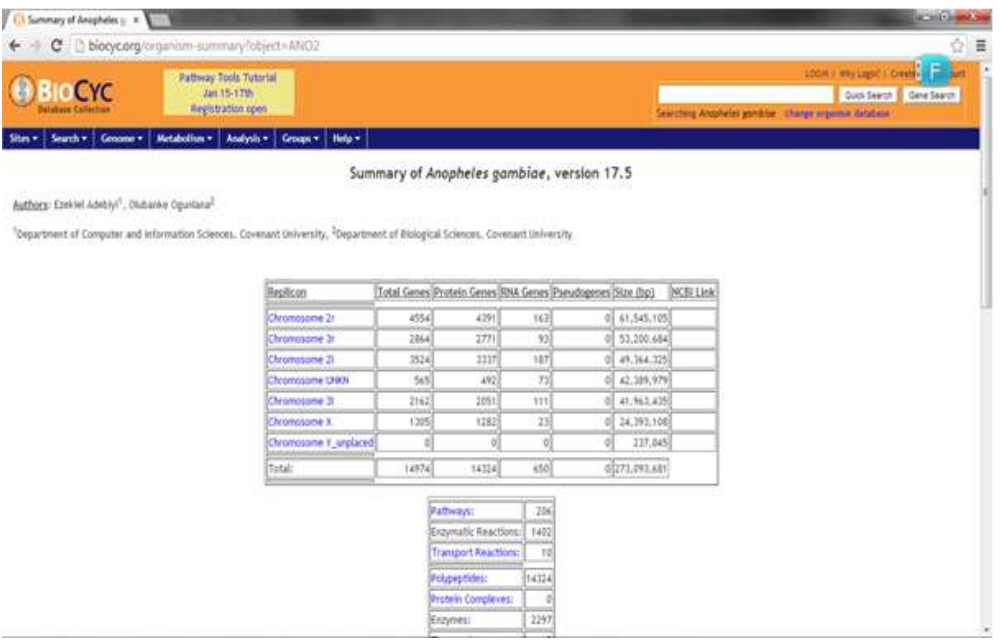

Figure 1.0: Home page of AnoCyc, version 1.0 on BioCyc [16]

The summary of this DB is depicted in the home page of AnoCyc, version 1.0 on BioCyc, while the metabolic overview of AnoCyc, version 1.0 is provided in Figure 2 [16]. There are 206 pathways, 135 (66%) of which are biosynthesis pathways for macromolecules (amino acids, carbohydrates, fatty acids/lipids and nucleosides/nucleotides), secondary metabolites, hormones, cellular structures, metabolic regulators, cofactors/prosthetic groups/electron carriers, aromatic compounds,

aminoacyl tRNA and amines/polyamines as well as other biosynthesis. 84 pathways (41%) are degradation/utilization/assimilation pathways. Others include 2 activation pathways for fatty acids and sulphate, 4 pathways for acid resistance, methylglyoxal, superoxide radical and glutathione-mediated detoxification, 24 super pathways and 2 transport pathways for copper and calcium. 16 pathways generate precursor metabolites and energy mainly the glycolysis, pentose phosphate pathways and TCA cycle while 4 constitute metabolic clusters of unrelated biochemical reactions.

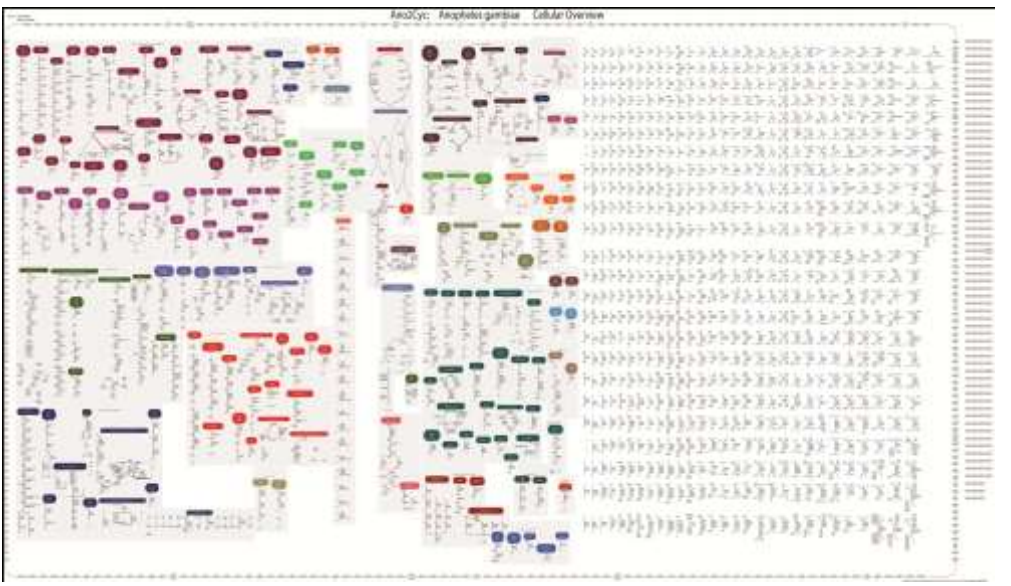

Figure 2.0: Overview of the metabolic map for *A. gambiae*. [16]

Each node here represents a metabolite, with the type of metabolite indicated by the shape of the node as indicated by the legend. The lines connecting the metabolites represent reactions. The enzymes, whose genes are identified in the genome, are showed by colored (shaded) shape.

This first version of AnoCyc formed the major backbone of this work, as the metabolic network was reconstructed and analyzed with the aim of elucidating some essential enzymatic reactions, pathways and even essential genes that may have an important role as potential insecticidal target. Further work and data from the re annotation of vector base, UNIPROT, and KEGG summarized on GenBank file was merged to the reconstructed network and was yet curated with the pathologic program, resulting in the second edition of the *Anopheles gambiae* PGDB, AnoCyc 1.1.

Implementation

Construction of a New PGDB for A. gambiae

Using the current annotated pathway genome database for *A. gambiae* is AnoCyc from BioCyc, a dedicated database for *Anopheles gambiae (*AnoCyc, Version 1.0), and to further update that current version to (AnoCyc, Version 1.0) the computational pipeline shown in Figure 3.0 was deployed, the

first version of AnoCyc 1.0 involved the creation of comprehensive metabolic reaction database, such a reaction database can be used to construct and analyze a metabolic network by linking pairs of reactions for which the product of one reaction is the substrate for other [9, 16], which was completely implemented in C programing language. However, AnoCyc 1.1 was implemented with GOPET and DomainSweep.

Updated GenBank files for *A. gambiae* were downloaded. Next, from KEGG, VectorBase, AnoBase and UNIPROT, informations like EC number and product (protein) definition for each gene were extracted. Then the two main standard annotation tools were engaged, namely; DomainSweep and GOPET [10, 11]. The two annotation tools are from the German Cancer Research Center (DKFZ) Heidelberg Unix Sequence Analysis Resources (HUSAR) based on a Convex port of the Unix version of the Wisconsin Package GCG (Genetics Computer Group) Inc., Madison, Wisconsin, USA). While GOPET was used to assign molecular function terms to cDNA or protein sequences utilizing Gene Ontology for annotation terms, it automatically predicts gene ontology terms for all proteins and maps protein databases by performing homology searches, using Support Vector Machines for the prediction and assignment of confidence values [10, 16]. DomainSweep identifies the domain architecture within a protein sequence and finds the correct functional assignments for an uncharacterized protein sequence. It employs different database search methods to scan a number of protein/domain family databases. It searches specifically for the most important protein family databases and in its output, domains are classified as "Significant" or "Putative" based on predefined rules such as database specific criteria of cutoff values or e-value thresholds, and so on. Domain hits are linked to the corresponding protein family database entries and are grouped together if they belong to the same InterPro family. Interpro - as an integrated resource - provides extensive domain annotations including direct access to the GO (gene ontology) classification system [11, 16]

The protein transcripts of *A. gambiae* genes are input to these tools to computationally gain insight into further annotation information for each gene. Also, several pieces of information like EC number and/ or product (protein) definition for each gene were extracted from the XML output files of these tools. The resulting reannotated GenBank file that was generated was inputted into the Pathway Tools to realize the new version for AnoCyc. Concluding the built of a PGDB (AnoCyc) for *A. gambiae* AgamP3 using the Pathologic program, the latest version, which is the second edition of the *A. gambiae* (PGDB) was generated by the Pathologic [12, 13, and 14], component of Pathway Tools software version 19.0 and MetaCyc version 19

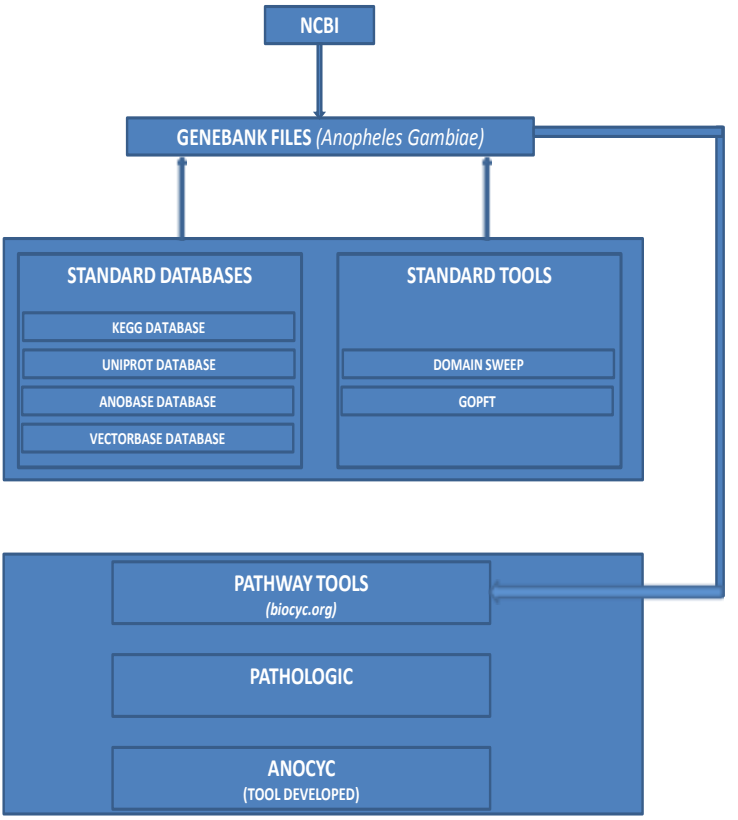

Figure 3.0: Schematic diagram of developmental stages of the Pathway/Genome Database (PGDB) for A*nopheles gambiae*. [16]

Results and Discussion

Using these methods, computational analysis of the *P. falciparum* network has revealed that curation helped to close gaps and link up dangling ends. Important information about the metabolism, transport, and genetic regulatory processes of this organism and more importantly, the pathologic program was all involved in its curation. The curated DB has her home page depicted in Figure 4.0, which is the second edition of AnoCyc, version 1.1

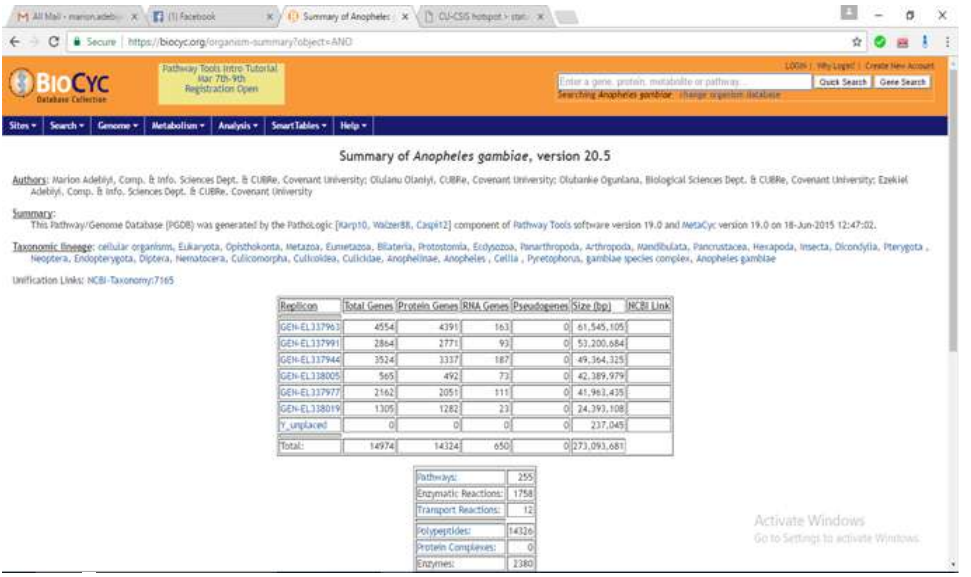

Figure 4.0: Home page of AnoCyc, version 1.1 on BioCyc

Same as in AnoCyc 1.0 (Figure 1.0), the summary of this DB as depicted in the home page of AnoCyc, version 1.1 (Figure 4.0), yet on BioCyc. The metabolic overview of AnoCyc 1.0 is provided in Figure 2 while AnoCyc 1.1 is explicitly represented in Figure 5.0. This knowledge base has a total size of about 273,093,681pb with 1237 compounds, 3684 Go terms, 650RNA genes, 75 transporters, 2380 enzymes without any protein complex, 14326 polypeptides, 12 transport reaction and 1758 enzymatic reactions.

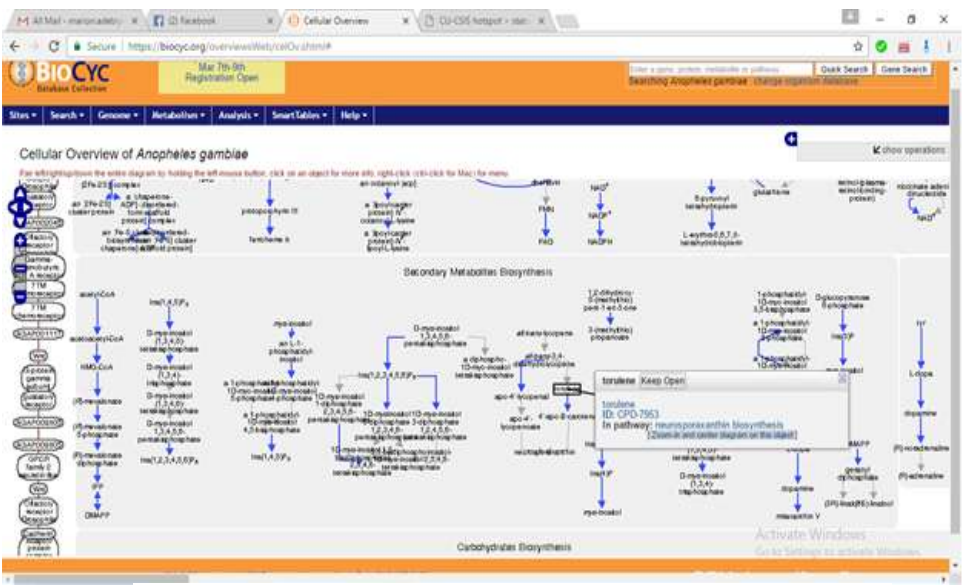

Figure 5.0: Cellular Overview of *Anopheles gambiae* from AnoCyc 1.1, BioCyc

Table 2: Legend of the *Anopheles gambiae* Cellular Overview

Object	Icon Shape
Amino Acids	Pointing up triangle
Carbohydrates	Square
Proteins	Diamond
Purines	Vertical ellipse
Pyrimidines	Horizontal ellipse
Cofactors	Pointing down triangle
tRNAS	Tee
Other	Circle
Phosphorylated	Filled

255 pathways were presented, 135 (66%) of which are biosynthesis pathways for macromolecules (amino acids, carbohydrates, fatty acids/lipids and nucleosides/nucleotides), secondary metabolites, hormones, cellular structures, metabolic regulators, cofactors/prosthetic groups/electron carriers, aromatic compounds, aminoacyl tRNA and amines/polyamines as well as other biosynthesis. 103 pathways (40%) are degradation/utilization/assimilation pathways. Others include 2 activation pathways for fatty acids and sulphate, 4 pathways for acid resistance, methylglyoxal, superoxide radical and glutathione-mediated detoxification, 42 super pathways and 1 transport pathways for copper and calcium. 18 pathways generate precursor metabolites and energy mainly the glycolysis, pentose phosphate pathways and TCA cycle while 9 constitute metabolic clusters of unrelated biochemical reactions. Also, 6 detoxification pathways, 7 macromolecule modification pathways, 16 nucleotides and nucleosides energy and 8 inorganic nutrients metabolism were presented in the DB. Table 2 has the legend details to help navigate around the cellular overview diagram.

Table 3: Summary of AnoCyc 1.0 (BioCyc, version 17.5) and AnoCyc 1.1 (BioCyc, version 20.5)

COMPONENT	AnoCyc 1.0 (BioCyc, version 17.5)	AnoCyc 1.1(BioCyc, version 20.5)
PATHWAYS	206	255
RNA genes	650	650
ENZYMATIC REACTION	1402	1758
SUPERPATHWAYS	24	42
TRANSPORTREACTIONS	10	12
POLYPEPTIDES	14324	14324
ENZYMES	2297	2380
TRANSPORTER	115	75
COMPOUNDS	958	1237
GO TERMS	3524	3684
TOTAL GENES	14974	14974

Using these methods, computational analysis of the *P. falciparum* network has revealed that curation helped to close gaps and link up dangling ends. The latest version, which is the second edition of the *Anopheles gambiae* Pathway/Genome Database (PGDB) was generated by the Pathologic [12, 13, and 14], component of Pathway Tools software version 19.0 and MetaCyc version 19.0.

Amazingly, the summary of some major components of AnoCyc 1.0 and 1.1 are detailed briefly in table 3, it enumerates the total number of genes, compounds and other important distinct chemical reactions curated from six various chromosomes of the organism. Closely-related variants of enzyme that showed that more than one form of enzyme catalyzes certain reactions were elucidated, also, very useful pathways and super pathways.

Conclusion

It is clear from the statistics in Table 1 that there is the need for further rigorous pursuit of the manual and automated curation of the biochemical metabolic network for *A. gambiae* [16], despite that we have very scanty data in available annotated databases. Yet a comprehensive and useful detail was generated from the limited information gotten about the organisms of interest in the various databases explored [15].

This work signifies the elucidation of information on Enzyme Classification (EC) numbers, biochemical pathways, compounds, reactions and Gene Ontology (GO) terms for the *Anopheles gambiae* proteins which has been excessively scanty and is now made available on the database. Experimental information on this are coming in too slowly and homology information on existing databases are not reliable. Presently, the closest vector to *A. gambiae,* with a standard pathway and genome database and other information is *drosophila melanogaster* and it is a very distant relative of the vector in consideration in this work.

References

1. Holt R. et al., (2002), "The Genome Sequence of the Malaria Mosquito *Anopheles gambiae*", Journal of Science, Vol. 298, pp.129 – 149.
2. Oakeshott, J. G., Irene Horne, Tara D. Sutherland and Robyn J. Russell (2003), "The genomics of insecticide resistance", Genome Biology, Vol. 4, pp.1-4.
3. Karp PD, Paley S, Romero P. The Pathway Tools software. Bioinformatics. 2002; 18 Suppl 1:S225-32.
4. Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N., Kunin, V., Lopez-Bigas, N., (2005). "Expansion of the BioCyc collection of pathway/genome databases to 160 genomes". Nucleic Acids Res. 33, pp. 6083- 6089.
5. Caspi Ron, Hartmut Foerster, Carol A. Fulcher, Pallavi Kaipa, Markus Krummenacker, Mario Latendresse, Suzanne Paley, Seung Y. Rhee, Alexander G. Shearer, Christophe Tissier, Thomas C. Walk, Peifen Zhang, and Peter D. Karp (2008), "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases", Nucleic Acids Research, Vol. 36, pp. 623-631.
6. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Paley S, Popescu L,

Pujar A, Shearer AG, Zhang P, Karp PD. (2010) "The MetaCyc database
7. pathway/genome databases," Nucleic Acids Research 38:D473-9.
8. Karp, P. (2001). "Many Genebank entries for complete microbial
9. Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27–30.
10. Adebiyi M**,** Fatumo S, Adebiyi E. (2013). "*In Silico* Study of the Complex Mechanisms of *Anopheles gambiae* Insecticide Resistance". Proceedings of the third joint meeting of ISCB and ASBCB and the fourth conference of the ASBCB on Bioinformatics of African Pathogens, Hosts and Vectors. Casablanca, Morocco from 11 – 16th March 2013.

10 Vinayagam A, del Val C, Schubert F, Eils R, Glatting KH, Suhai S, Koenig R.GOPET: a tool for automated predictions of Gene Ontology terms.BMC Bioinformatics. 2006 Mar 20;7:161.

11. Del Val C, Ernst P, Falkenhahn M, Fladerer C, Glatting KH, Suhai S, Hotz-Wagenblatt A.ProtSweep, 2Dsweep and DomainSweep: protein analysis suite at DKFZ.Nucleic Acids Res. 2007 Jul;35(Web Server issue):W444-50. Epub 2007 May 25.
12. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD (2012**).** "The MetaCyc of metabolic pathways and enzymes and the BioCyc collection of genomes violate the Genbank standard". Comparative and Functional Genomics, Vol. 2. pp. 25-27.
database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases." Nucleic Acids Res 40(D1); D742-D753. PMID: 22102576.
13. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, Altman T, Paulsen I, Keseler IM, Caspi R (2010). **"Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology."** Brief Bioinform 11(1);40-79. PMID: 19955237
14. Walzer Y, (1988). "Female urinary incontinence. Discerning the exact cause." Postgrad Med 83(7); 78-80, 86-8. PMID: 3368424.

15 Rohwer, M. and Snaep, L. "Structural Properties of Metabolic Networks: Implications for Evolution and Modeling of Metabolism". Stellenbosch University Press, Stellenbosch, pp. 79–85.

16 Adebiyi, Marion O., *Computational analysis of Anopheles gambiae metabolism to facilitate insecticidal target and complex resistance mechanism discovery*. PhD Dissertation, Covenant University, 2014.
http://theses.covenantuniversity.edu.ng/handle/123456789/953

Web Refference

http://biocyc.org/otherpgdbs.shtml
http://www.map.ox.ac.uk/explore/malaria-vectors

www.anobase.org
(www.genome.jp/kegg/)
www.biocyc.org
www.ebi.ac.uk
www.ensembl.org
www.pdb.org , www.mosquitoes.com
www.rcsb.org
http://biocyc.org/otherpgdbs.shtml