SEARCHING THE GENE ONTOLOGY TERMS USING SEMANTIC SIMILARITY MEASURE

Muhamad Razib Othman, Safaai Deris, and Rosli Md. Illias

***Abstract*—**The most important property of the Gene Ontology is the terms. These control vocabularies are defined to provide consistent descriptions of gene products that are shareable and computationally accessible by humans, software agent, or other machine-readable meta-data. Each term is associated with information such as definition, synonyms, database references, amino acid sequences, and relationships to other terms. This information has made the Gene Ontology broadly applied in microarray and proteomic analysis. However, the process of searching the term is still carried out using traditional approach which is based on keyword matching. The weaknesses of this approach are: ignoring semantic relationships between terms, and highly depending on a specialist to find similar terms. Therefore, semantic similarity measure is used to compute similitude strength between terms and computational results are presented.

***Keywords*—**Gene Ontology, ontology, search, semantic similarity measure.

I. Introduction

THE Gene Ontology (GO) [1] is a biological ontology maintained by the GO Consortium which is located at www.geneontology.org. The project attempts to provide a consistent term to describe gene and gene product in any organism found in heterogeneous databases. GO plays an important role in searching biological information and annotating proteins or genomes. Some examples of GO applications include prediction of functional modules [2], microarray analysis [3], prediction of protein-protein interactions [4], and proteomics analysis [5].

The amount of available GO terms has grown enormously and become more demanded in the last few years. A total number of 628 articles was related to the GO since 1998 as shown in Fig. 1. Although tools for searching the GO terms such as AmiGO (www.godatabase.org), GenNav (mor.nlm.nih. gov/perl/gennav.pl), QuickGO (www.ebi.ac.uk/ego/), and MGI GO Browser (www.informatics.jax.org/searches/GO_form. shtml) are publicly available, these search engines respond to user keyword queries by retrieving relevant GO terms based on word matching or Boolean rules.

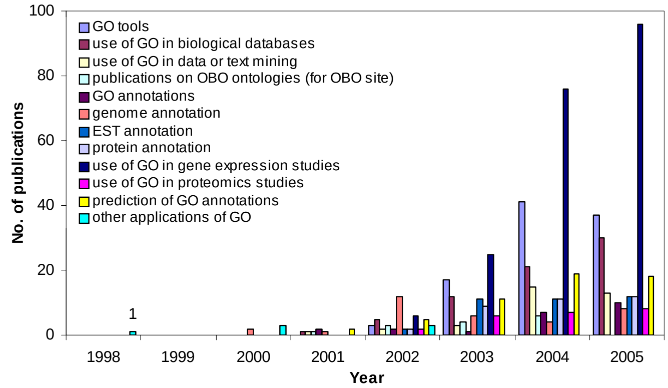

Fig. 1 Gene Ontology bibliography

In response to this scenario, an approach to search the GO terms is proposed using semantic similarity measure to determine the similitude strength of two terms organized in the GO graph (see Section 2 for formal definition). This semantic similarity measure (see Section 4) is a hybrid approach by combining information content and conceptual distance. The information content will compute the amount of information the GO terms share in common. On the other hand, the conceptual distance will calculate the depth and the local network density of the GO term. Furthermore, this study will accommodate biologists as well as alignment tools such as BLAST (www.ncbi.nlm.nih.gov/BLAST/), CLUSTALW (www.ebi.ac. uk/clustalw/), and SIM (www.expasy.ch/tools/ sim-prot.html) to reduce the processing time of discovering similar sequences. As a matter of fact, Lord et al. [6] has presented results showing the correlation between semantic similarity and sequence similarity.

The rest of the paper is organized as follows. Section 2 begins with the problem description of ontology search. Section 3 gives a review of related work in search of the GO terms and semantic similarity measure. Section 4 discusses the technical description of the proposed semantic similarity measure. Section 5 presents experimental results and is followed by discussion of the results in Section 6.

II. Problem Description

Ontology is a description of concepts in a domain and the relationships between the concepts. Ontology can be represented as a directed graph. The ontology graph comprises the concepts including the descriptions as nodes and semantic relationships as edges. Recently, there has been growing development of ontology in the bioinformatics field such as

Muhamad Razib Othman is with the Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, MALAYSIA (corresponding author; phone: 07-5532358; fax: 07-5565044; e-mail: razib@ fsksm.utm.my).

Safaai Deris is with the School of Graduate Studies, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, MALAYSIA (e-mail: safaai@fsksm.utm.my).

Rosli Md. Illias is with the Faculty of Chemical and Natural Resources Enginneering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, MALAYSIA (e-mail: r-rosli@utm.my).

Sequence Ontology [7], Cell Ontology [8], Chemical Ontology [9], Multiple Alignment Ontology [10], and Biodynamic Ontology [11]. By contrast, the "ontology search" which is referring to the activity of retrieving concepts in the ontology graph is not accurately performed by the traditional search engines that are based on keywords. These search engines neglect the semantic relationships of the search concepts and only consider those concepts as character strings. Thence, mechanism to measure the similarity between concepts in the ontology graph is required to reduce dependency of specialists of a certain domain to input relevant concepts as query words.

Given a GO graph $G = \{V, E\}$ that is structured as a Directed Acyclic Graph (DAG). V is a finite non-empty set of nodes representing GO terms and E is a finite set of pairs of nodes representing relationships between GO terms. Each pair in E is an arc of G. The GO terms can have more than one parent, as well as multiple children. The GO terms are linked by two relationships, the "is-a" relationships ("intracellular organelle", GO:0043229 and "membrane-bound organelle", GO:0043227 are parent of "intracellular membrane-bound organelle", GO:0043231) and the "part-of" relationships ("chloroplast stroma", GO:0009570 is part of "chloroplast", GO:0009507).

Searching the GO graph to retrieve semantically similar terms is a NP-complete problem. This is due to the size of the search space of the DAG as $g(k)$ is astronomical and vary between:

$$2^{\frac{k(k-1)}{2}} \leq g(k) \leq 3^{\frac{k(k-1)}{2}} \quad (1)$$

where k is the number of nodes in the GO graph. To search the GO graph, the following research problems need to be figured out:

1) What is the most suitable search algorithm for finding feasible solution that offers reasonable amount of time to this NP-complete problem?
2) What is the precise criterion to this ontology search problem for quantifying the semantic similarity between GO terms?

Focus of this paper is to solve the second problem using semantic similarity measure in order to assist search techniques to perform batch retrievals that have the ability to search one term towards all terms in the GO graph.

III. Review of Related Work

Several GO browsers have been developed to provide text searching over the GO terms and associated information such as definition, synonyms, lineage, cross-references, and gene products annotated to them. These browsers also have graphical view of the hierarchy of the target terms. A comprehensive overview with links to respective addresses can be accessed at www.geneontology.org/GO.tools.browsers.shtml. Among these tools are:

3) AmiGO, a GO browser developed by the GO Consortium. The keyword-based search is executed either by exact or approximate match over the term accession number, name, or synonyms. This tool also allows a user to use gene product or protein sequence as search input.
4) GenNav, a GO browser that uses string matching method namely exact or approximate match that responds to a given term or gene product. GenNav is maintained by the United States National Library of Medicine.
5) QuickGO, a GO browser that allows user to retrieve the GO terms by exact or wildcard search over the term accession number, name, synonyms, definitions, or comments. This fast web-based GO browser can be found at the European Bioinformatics Institute website.
6) MGI GO Browser, a GO browser developed by the Mouse Genome Informatics that perform string matching by requiring users to enter partial term name or full term accession number.
7) EP GO Browser, a GO browser that carries out the exact or contains match to the term accession number or name entered by the user. This browser is built into an expression profiler developed by the European Bioinformatics Institute.

Lately, semantic similarity measure has been introduced in many areas related to natural language processing and information retrieval. For example, this measure has been applied in the ontology integration [14], environmental modeling [15], computational linguistics [16], and bioinformatics [17]. Semantic similarity measure has the capability to improve the precision and recall of information retrieval by discovering the correlation between concepts. This is done by computing the relatedness between concepts either by estimating the distance or the amount of information in the commonality of the two concepts being compared. Most popular mechanisms used to calculate the semantic similarity between concepts are founded by [18]–[20]. The comparison in [21] shows that Jiang and Conrath's semantic similarity provides the best results, and it is used as a main reference in this study.

IV. The Proposed Semantic Similarity Measure

A. Information Content

The information content is calculated according to "association", a source showing information that is shared among the GO terms. The association is a table which stores annotations that basically provide a link between a gene product and a GO term with an evidence code. For example, a gene product "dynein, axonemal, heavy chain 11" (Dnahc11) is associated to several GO terms such as "determination of left/right symmetry" (GO:0007368) with an evidence code of IMP (Inferred from Mutant Phenotype), "axonemal dynein complex" (GO:0005858) with an evidence code of IDA (Inferred from Direct Assay), and "mitochondrial inner membrane" (GO:0005743) with an evidence code of RCA (inferred from Reviewed Computational Analysis). The information content of the GO term $IC(v)$ is given by the following equation:

$$IC(v) = -\log(P(v)) \quad (2)$$

where $P(v)$ is the probability for the occurrence of a GO term v in the association. This probability can be computed using maximum likelihood estimation as below:

$$P(v) = \frac{freq(v)}{N} \quad (3)$$

where N is the total number of occurrences in the association and $freq(v)$ is the number of times that the GO term v and all its descendants occur in the association. The frequency of the GO term v is given as follows:

$$freq(v) = \sum_{v \in descendants(v_i)} occur(v_i) \quad (4)$$

where $descendants(v)$ is a function that returns the set of GO terms that are the descendants of the GO term v. Note that, if a GO term v_a is an ancestor of a GO term v_b, then $freq(v_a) \geq freq(v_b)$ since the GO term v_a subsumes the GO term v_b and all its descendants. Therefore, $P(v)$ is larger when the GO term v is closer to the root term v_0 and $IC(v_a) \leq IC(v_b)$.

B. Conceptual Distance

The conceptual distance of the GO term is measured by the depth and the local network density factors. The depth is related to the distance of the GO term in the hierarchy of the GO graph. The local network density is associated to the number of children that span out from the GO term. The depth of the GO term $D(v)$ is represented as below:

$$D(v) = \left(\frac{d(v)+1}{d(v)} \right)^{\alpha} \quad (5)$$

where $d(v)$ is the level of the GO term v in the GO graph. The $d(v)$ of the root term v_0 is 1 and increases as the GO term altitude moves downward in the hierarchy. The parameter α controls the degree of how much the depth factor contributes in (5) and $\alpha \geq 0$.

The local network density of the GO term $E(v)$ is defined as follows:

$$E(v) = \left((1-\beta) \times \frac{\overline{E}}{e(v)} \right) + \beta \quad (6)$$

where $e(v)$ is the number of edges that begin from the GO term v and $\overline{E}$ is the number of edges divided by the number of GO terms that exist in the GO graph. The parameter β controls the degree of how much the local network density factor contributes in (6) and $0 \leq \beta \leq 1$.

The parameters α and β become less important when α approaches 0 and β approaches 1 since $D(v)$ and $E(v)$ will approach 1 respectively. Furthermore, (5) and (6) are equivalent when $\alpha = 0$ and $\beta = 1$.

C. The Hybrid Approach

The hybrid approach is derived from the conceptual distance notion and integrates the information content as a decision factor. Given a sequence of GO terms v_a, …, v_n representing the path from GO term v_a to v_n with length n. The hybrid approach calculates the semantic distance between GO term v_a and v_n by the given formula:

$$dist(v_a, v_n) = \sum_{i=0}^{n-1} D(v_i) \times E(v_i) \times \left(IC(v_{i+1}) - IC(v_i) \right) \quad (7)$$

where $dist(v_a, v_n)$ is the summation of edge weights along the shortest path that link v_a with v_n. Thus, the semantic distance between GO term v_m to v_n is quantified as follows:

$$dist(v_m, v_n) = dist(v_a, v_m) + dist(v_a, v_n) \quad (8)$$

where GO term v_a is the closest shared ancestor of GO term v_m and v_n. Since the semantic distance is based on the difference between the information content, the normalization of the semantic distance is given by:

$$dist_{norm}(v_m, v_n) = \min\{1, \frac{dist(v_m, v_n)}{\max\{IC(v)\}}\} \quad (9)$$

Therefore, the semantic similarity measure between GO term v_m to v_n is calculated by converting the semantic distance as follows:

$$SSM(v_m, v_n) = 1 - dist_{norm}(v_m, v_n) \quad (10)$$

Note that, $0 \leq SSM(v_m, v_n) \leq 1$ because $0 \leq dist_{norm}(v_m, v_n) \leq 1$.

V. Experimental Results

The proposed semantic similarity measure (A) has been tested using GO data from [22]. The results are compared with other semantic similarity measures proposed by Jiang and Conrath (B), Lin (C), and Resnik (D). The results in Table 1 shows an increase of similarity percentage for the proposed semantic similarity measure.

In order to evaluate the applicability of the proposed semantic similarity measure in searching the GO terms, its formula is added into genetic algorithm during the creation of population and calculation of fitness value [23]. The parameters used to run the genetic algorithm are shown in Table 2. The computer used is HP d530 with Pentium 4 processor 2.8 GHz, 512 MB RAM, and 100 Mbps NIC running under Fedora Core 2.

The stability of the proposed semantic similarity measure can be seen in Table 3 and Fig. 2, where results of 3 separate runs are compared. The convergence appeared as early as after 430 generations. The optimal value of the fitness function is in

TABLE I
COMPARISON WITH OTHER SEMANTIC SIMILARITY MEASURES

Term Accession Number	Term Name	A	B	C	D
GO:0005575	cellular component	18.2	18.2	0.0	0.0
GO:0005622	intra cellular	15.8	15.3	9.1	1.1
GO:0005623	cell	19.2	18.6	11.0	1.1
GO:0005737	cytoplasm	13.6	12.9	7.9	1.1
GO:0009536	plastid	9.1	8.8	5.5	1.1
GO:0016020	membrane	25.2	23.4	18.4	6.0
GO:0019866	organelle inner membrane	100.0	100.0	100.0	19.0
GO:0019867	outer membrane	7.8	7.2	6.9	6.0
GO:0031090	organelle membrane	12.4	10.1	8.1	6.0
GO:0043226	organelle	13.3	13.2	0.0	0.0
GO:0043227	membrane-bound organelle	12.7	12.5	0.0	0.0
GO:0043229	intra cellular organelle	14.5	13.9	8.4	1.1
GO:0043231	intra cellular membrane-bound organelle	13.8	13.2	8.0	1.1

Notes: Terms pair with "organelle inner membrane" (GO:0019866) and results are in similarity percentage.

TABLE 2
PARAMETERS OF GENETIC ALGORITHM

Item	Parameter
Number of population	100
Number of generation	1000
Crossover probability	0.8
Mutation probability	0.01
Size of chromosome	19589
Replacement percentage	0.5
Type of crossover	Two-point crossover
Type of mutation	Swap mutation
Type of genetic algorithm	Steady-state genetic algorithm
Scaling	Sigma truncation scaling
Fitness function	Maximizing preferences

TABLE 3
RESULTS OF THREE RUNS

Items	Run 1	Run 2	Run 3
Processing time (seconds)	17	23	11
No. of generation to converge	590	630	430
Maximum value of fitness function	1616.1	1616.1	1610.7

the interval 1610.7 to 1616.1. The time taken is varied from 11 seconds to 23 seconds.

VI. DISCUSSION

In this paper, an approach for measuring semantic similarity between GO terms is presented. The proposed measure is a combined approach that inherits the edge-based approach of the edge counting scheme, which is enhanced by the node-based approach of information content calculation. When tested, the proposed measure outperforms other semantic similarity measures. By combining with search technique, specifically genetic algorithm, the experimental results show that the proposed measure is effective, stable, and thus, it required reasonable amount of execution time. Possible directions for further research would be to include evidence codes during the calculation of the information content. In this way, the degree of information the GO terms share in common will be more accurate and correspond to evidence such as genetic interaction, sequence similarity, expression pattern, mutant phenotype and others that support the GO annotation.

ACKNOWLEDGMENT

This material is based upon work supported by the Malaysian Ministry of Science, Technology, and Innovation (MOSTI) in part under Intensification of Research in Priority Areas (IRPA) grants (project no. 04-02-06-0057-EA001 and 04-02-06-10050-EAR) and in part under Short Term Research (STR) grant (project no. 75162).

REFERENCES

[1] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G.. Sherlock, "Gene ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25-29, May 2000.

[2] H. Wu, Z. Su, F. Mao, V. Olman, and Y. Xu, "Prediction of functional modules based on comparative genome analysis and gene ontology application," *Nucleic Acids Res.*, vol. 33, no. 9, pp. 2822-2837, May 2005.

[3] J.A. Young, Q.L. Fivelman, P.L Blair, P. de la Vega, K.G. Le Roch, Y. Zhou, D.J. Carucci, D.A Baker, and E.A. Winzeler, "The plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification," *Mol. Biochem. Parasitol.*, vol. 143, no. 1, pp. 67-79, Sep. 2005.

[4] J. Espadaler, O. Romero-Isart, R.M. Jackson, and B. Oliva, "Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships," *Bioinformatics*, vol. 21, no. 16, pp. 3360-3368, Aug. 2005.

[5] S.M. Hauck, S. Schoeffmann, C.A. Deeg, C.J. Gloeckner, M.S. Lange, and M. Ueffing, "Proteomic analysis of the porcine interphotoreceptor matrix," *Proteomics*, vol. 5, no. 14, pp. 3623-3636, Sep. 2005.

[6] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble, "Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275-1283, Jul. 2003.

[7] K. Eilbeck, S.E. Lewis, C.J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner, "The sequence ontology: a tool for the unification of genome annotations," *Genome Biol.*, vol. 6, no. 5, rec. R44, Apr. 2005.

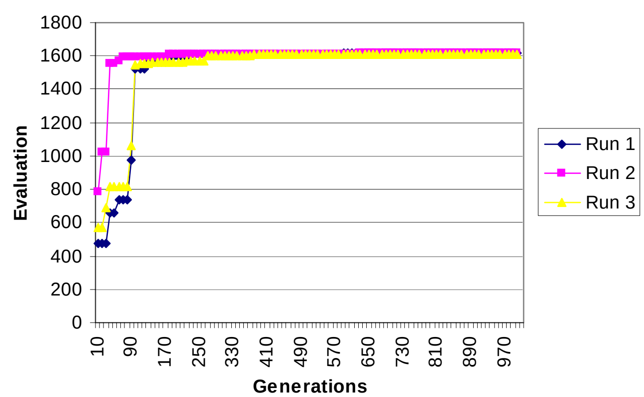

Fig. 2 Evolution of 3 runs

[8] J. Bard, S.Y. Rhee, and M. Ashburner, "An ontology for cell types," *Genome Biol.*, vol. 6, no. 2, rec. R21, Jan. 2005.

[9] H.J. Feldman, M. Dumontier, S. Ling, N. Haider, and C.W. Hogue, "CO: a chemical ontology for identification of functional groups and semantic comparison of small molecules," *FEBS Lett.*, vol. 579, no. 21, pp. 4685-4691, Aug. 2005.

[10] J.D. Thompson, S.R. Holbrook, K. Katoh, P. Koehl, D. Moras, E. Westhof, and O. Poch, "MAO: a multiple alignment ontology for nucleic acid and protein sequences," *Nucleic Acids Res.*, vol. 33, no. 13, pp. 4164-4171, Jul. 2005.

[11] P. Grenon, B. Smith, and L. Goldberg, "Biodynamic ontology: applying BFO in the biomedical domain," *Stud. Health Technol. Inform.*, vol. 102, pp. 20-38, Apr. 2004.

[12] H. Liu, Z. Hu, and C.H. Wu, "DynGO: a tool for browsing and mining gene ontology and its associations," *BMC Bioinformatics*, vol. 6, rec. 201, Aug. 2005.

[13] F. Couto, M. Silva, and P. Coutinho, "Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors," presented at the 14th ACM Conf. Information and Knowledge Management, Bremen, Germany, Oct. 31 - Nov. 5, 2005.

[14] M.A. Rodriguez and M.J. Egenhofer, "Determining semantic similarity among entity classes from different ontologies," *IEEE Trans. Knowledge and Data Engineering*, vol. 15, no. 2, pp. 442-456, Mar. 2003.

[15] C.-C. Feng and D.M. Flewelling, "Assessment of semantic similarity between land use/land cover classification systems," *Computers, Environment, and Urban Systems*, vol. 28, no. 3, pp. 229-246, May 2004.

[16] G. Vigliocco, D.P. Vinson, and S. Siri, "Semantic similarity and grammatical class in naming actions," *Cognition*, vol. 94, no. 3, pp. B91-B100, Jan. 2005.

[17] O. Steichen, C.D. Le Bozec, M. Thieu, E. Zapletal, and M.C. Jaulent, "Computation of semantic similarity within an ontology of breast pathology to assist inter-observer consensus," *Computers in Biology and Medicine*, to be published.

[18] D. Lin, "An information-theoretic definition of similarity," in *Proc. 15th Int. Conf. Machine Learning*, Madison, WI, 1998, pp. 296-304.

[19] J.J. Jiang and D.W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. 1998 Int. Conf. Research in Computational Linguistics*, pp. 19-33.

[20] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. 14th Int. Joint Conf. Artificial Intelligence*, Montreal, Canada, 1995, pp. 448-453.

[21] A. Budanitsky and G. Hirst, "Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures," presented at the 2nd Meeting North American Chapter of the Association for Computational Linguistics, Pittsburgh, PA, Jun. 2-7, 2001.

[22] M.R. Othman, S. Deris, R.M. Illias, Z. Zakaria, and M.S. Mohamad, "Automatic clustering of gene ontology by genetic algorithm," To appear in *Int. J. Information Technology*.

[23] M.R. Othman, S. Deris, R.M. Illias, and Z. Zakaria, "Semantic similarity measure over genetic algorithm: an approach for searching the gene ontology terms," *Int. J. Computational Intelligence*, submitted for publication.

Muhamad Razib Othman is a doctoral candidate at the Faculty of Computer Science and Information System, the Universiti Teknologi Malaysia. He received the BSc and MSc degrees in Computer Science both from the Universiti Teknologi Malaysia. Currently, he is working for his PhD in Computational Biology. He also has interest in artificial intelligence, software agent, parallel computing, and web semantics. In March 2005 he has been awarded the Young Researcher bestowed by the Malaysian Association of Research Scientists (MARS). One of his inventions, a software product named *2D Engineering Drawing Extractor,* has recently won 5 awards including the Best Invention of the Pacific Rim at the 21st Invention and New Product Exposition (INPEX) held in Pittsburgh, USA.

Safaai Deris is a Professor of Artificial Intelligence and Software Engineering at the Faculty of Computer Science and Information System, Deputy Dean at the School of Graduate Studies, and Director of Laboratory of Artificial Intelligence and Bioinformatics at the Universiti Teknologi Malaysia. Safaai holds DEng in Computer and System Sciences and MEng in Industrial Engineering both from the Osaka Prefecture University, Japan. His recent academic interests include the application and development of intelligent techniques in planning, scheduling, and bioinformatics.

Rosli Md. Illias is an Associate Professor at the Faculty of Chemical and Natural Resources Engineering at the Universiti Teknologi Malaysia. Rosli received his PhD degree in Molecular Biology from the Edinburgh University, UK and BSc degree in Microbiology from the Universiti Kebangsaan Malaysia. His research interests are in the areas of microbial technology, molecular enzymology, and molecular genetics.