TM-0697

The Fifth Generation Computer Technology and Biological Sequencing

by
S. Uchida & K. Yoshida

March, 1989

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03) 456-3191~5
Telex ICOT J32964

Institute for New Generation Computer Technology

The Fifth Generation Computer Technology and Biological Sequencing

Shunichi Uchida Kaoru Yoshida

Institute for New Generation Computer Technology (ICOT)

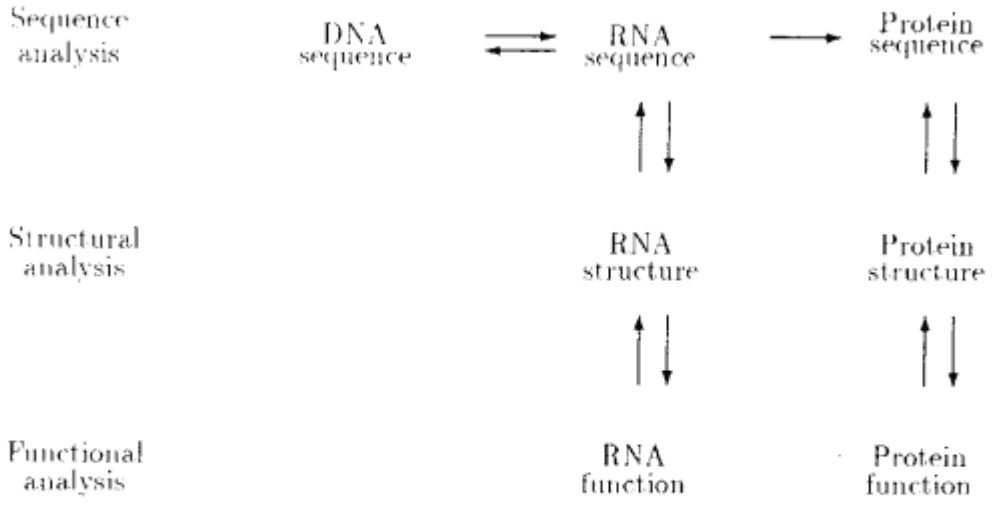
Mita-Kokusai Building 21F., 4-28, Mita 1, Minato-ku, Tokyo 108

1 Introduction

This informal paper intends to answer the questions given to the participants of the work shop on advanced computer technologies and biological sequencing held at the Argonne National Laboratory. It also includes some comments on the applicability of advanced computer technology, especially, the fifth generation computer technology, to biological sequencing and functional analysis of DNA/RNA/proteins. The contents of the answers and comments are author's personal opinions and hopes as researchers of advanced computer technology.

2 Problem Solving in Molecular Biology

In this section, we make sure of our understanding of the problem solving of molecular biology and overview our comments from the 5G technology's point of views. That is what are or will be the problems to be solved and the limiting factors for now, which ould require advanced computer technologies including the 5G computer technology.



1. **High parallelism of the homology searching:** The entire problem solving of molecular biology may contain different kinds and levels of high degree of parallelism in different parts of the problem, e.g.

- Most of the algorithms in use for the sequence analysis are based on the multivariate analysis which is supposed to require enormous computation power for solution.

  → Since the contained parallelism is uniform and regular, SIMD computer architectures may be appropriate.

- The structural analysis will not stay at the RNA or protein level. In the future, it will be connected to the cell and organism levels. In these structural analyses, dynamic simulation for both quantative and qualitative reasoning will take a large part.

  → Fine grained highly parallel MIMD computing environments will be necessary.

- For the functional analysis, there are numerous unknown things

  → Some expert system technology will be needed for heuristic reasoning.

2. **Exponentially growing databases:**
   The amount of data entries of databases is increasing exponentially year by year. Many of the existing databases are build up using the conventional relational database models which will result in the problems on database overflow and impractically low performance as the data amount increases.

   → The databases should be built up adopting new database models, such as nested relational models and object-oriented databases and so on.

3. **Integrated software system:** It seems that there have not so many software systems developped for molecular biology. A large scale of different kinds of software will need to be developed from now.

   Furthermore, many of the existing software systems are for a partial function and there are almost no integrated system which enables to reason from some sequence to a function or vice versa.

   → For ease of development, maintenance and extension, the development of the software should be done in high level languages.

4. **Friendly User Interface:** Looking at the existing molecular biological software systems, it is hard to say their user interface is simple and friendly to the end users, biologists. In other words, they are computer oriented rather than user-oriented.

   → For the system to be accepted naturally by biolosists, it is necessary to develop software systems with friudly user interface.

5. **Up-to-date research environment:**
   The research progress is so rapid in molecular biology.

   → For biologists to come up, share and exchange the newest information time to time, it will be necessary to build up a world-wide computer network environment.

3 Applicable 5G Technology

We have been proceeding the fifth generation computer technology project at ICOT, with logic programming as its central notion.

Among the research and development results which have been brought out in the project, we introduce those which may be applicable for and contribute to the above problem solving in molecular biology.

1. **High level programming languages**

   - Logic programming languages ... KL0 (Prolog-like) [Yokoi84], ESP [Chikayama84], CIL [Mukai88], GHC [Ueda85], KL1(GHC+α - β) [Chikayama88]
   - Constraint programming languages ... CAL [Aiba88]
   - Object-oriented programming languages ... ESP, A'UM [Yoshida88]
   - Parallel (or concurrent) programming languages ... GHC, KL1, A'UM

2. **System software development technology:**

   - Sequential operating system: SIMPOS which is an operating system written in ESP/KL0 to run ESP/KL0 programs on PSI and PSI-II machines [Yokoi84].
   - Parallel operating system: PIMOS which is an operating system written in KL1 to run KL1 programs on Multi-PSI and PIM machines [Chikayama88].

3. **Inference machine architectures:**

   - Sequential inference machines:
     - PSI [Nakajima86] and PSI-II [Nakashima87] tuned for KL0 execution.
     - CHI [Nakazaki85] tuned for Prolog execution
   - Parallel inference machines:
     - Multi-PSI [Taki86] designed for parallel software research
     - PIM [Goto88, Uchida88A] tuned for KL1 execution

4. **Distributed network environments:**

   At present, about 200 copies PSI and PSI-II machines are connected together via local and global area networks.

   - Workstations: PSI, PSI-II (besides SUN)
   - Central machines: Multi-PSI, PIM (besides DEC-2065, VAX750/780, Symmetry)
   - Network protocols: PSI-net (besides DEC-net, TCP/IP)

5. **Knowledgebase and database management systems:**

   - Nested relational and object-oriented knowledge base system: KAPPA [Yokota88]

6. **Expert systems:**

   - Expert shells/systems: [Fujii88]
     - Constraint-based probelem solving and object modeling
     - Hypothetical reasoning
     - Distributed cooperative problem solving
     - Qualitative reasoning

Knowlege acquisition support system

7. **Natural language understanding:**

   - Discourse understanding systems: DUALS (-I, -II, -III) [Uchida88B].

4 Questions and Answers

In this section, we answer to the given questions, focusing on what of the above mentioned 5G technologies will be available in practical use in 1991-1994.

4.1 General Hardware Trends

1. **Speeds of Processors:**
   General purpose microprocessors will attain 10-15 MIPS by 1991. They will contain 1.0-1.5M transistors per chip. Their clock frequency will be 25-33 MHz. Around 1994, they may attain 20-30 MIPS. Large mainframes will attain 100-200 MIPS. However, they will be a multi-processors containing 4 to 8 CPUs.

2. **Memory Technology:**
   Most workstations will have 50-100 MB for main memories in 1991 using 4M bit chips. Some workstations will have more memories. For example, a logic programming workstation of ICOT, PSI machine has 80 MB as a standard configuration. It can be extended to 160 MB.

   Mainframes may have 2 to 10 GB in 1991-1994.

3. **Disk Technology:**
   Most workstations will have disk of 200-400 MB as a standard configuration. Some expensive workstations may have larger disks, for example, more than 1 GB in 1991-1994.

   PSI machine has a 200 MB disk as a standard. It can be extended to 800 MB.

4.2 General Software Trends

1. **Computing Environments:**
   Unix will have more merits than DEC VMS or IBM MVS because hardware systems for Unix will cover from a tiny workstation to a powerful multiprocessor. (However, Unix can not support real parallel processing.)

   Unix also provides a distributed network environment which contains many personal workstations and a few big center computers. Environments of this type will be one of the best solutions in the near future.

   In these environment, the workstations will be used for small to medium scale static analysis like sequence matching and functional analysis. The center computers will be used for simulations of dynamic molecular structures or three dimensional graphics.

   Conventional large mainframes will be used in the first stage of biological research. However, complex problems like analysis of 3D structures and functional analysis of

proteins will exceed potential capabilities of conventional general purpose computers. These problems needs not only sufficient computing power but also high level knowledge processing capability like expert systems. New parallel computer and AI software technology can contribute more to these big and complex problems.

2. **Language Developments:**
It is obvious that higher level programming languages, such as functional and logic programming languages, will bring biologists much more merits than conventional procedural languages. Of course, C or C++ is better than Fortran, but still too low. Functional or logic programming languages are much better than these procedural languages in programming, debugging, maintaining and extending, if their hardware supports are sufficient and realize high performance per cost.

Presently, not so many software systems are existing for biological research. Thus, the software development must be guided to use high level languages which will make full use of advanced computer technology in the future.

Furthermore, molecular biological researches cannot stay at the level of DNA/RNA/proteins. In the future, they will be connected or expanded to researches at the level of cells, neurons and organisms. Including the researches on 3D molecular dynamics, simulation at each of the above levels will be one of the largest applications in this field. There are very few or almost no good programming languages and execution environments to simulate such a great number of fine grained communicating processes.

ICOT's concurrent logic programming languages and execution environments will greatly contribute to this kind of problem solving.

3. **Database Technology:**
Main databases being accumulated in this field seem to consists of a lot of sequences of DNA fragments. Currently, data processing applied to these databases do not seem to be so complex. Then conventional database technology using conventional computers can be used effectively.

However, if biologists want to handle more sophisticated data structures like rules in which relations between functions of proteins and their 3D structures are described, advanced database technology will be necessary.

4.3 Special-purpose Hardware and General-purpose Hardware

1. **Performance improvements:**
As far as simple but heavy problems like a homology search are concerned, special purpose hardware will attain better performance than general-purpose hardware. Parallel computers like pipe-lined processors for scientific computation greatly contribute these problems. If the homology search has to be combined with intelligent processing using a variety of knowledge like advanced expert systems, MIMD type parallel computers will be important.

2. **Cost/benefit:**
For the homology search, special-purpose hardware will be worth building. However, this type of hardware has to compete with scientific computers which can have a wider market. Then, the hardware building must be planned carefully.

4.3.1 Developments in Multiprocessing Capabilities

1. **Shared-memory MIMD machines:**
   The number of processors of a shared memory MIMD machines will be less than 20 in 1991-1991. (If the speed of processors is slow the number could be larger.) Its hardware price may be $200,000 to $500,000. Its performance may be around 100 MIPS at maximum but depends on algorithms and characteristics of the problem to be solved.

2. **MIMD computers without shared memory:**
   As the number of processors which can be connected to a shared memory is limited, a large scale MIMD computer has to use a loosely coupled network.

   For such a large scale MIMD computer, a shared memory MIMD machine can be used as a node of a loosely coupled network. This organization can combine merits of these two types of MIMD machines. One of ICOT's parallel inference machines (PIM-p) is of this type.

   It is likely that machines of this type will appear in the market soon. However, software systems for biological research have to be matured before they can be effectively used on the machines.

3. **SIMD machines:**
   Since the homology search is substantially simple, mechanical and regular in its algorithm, SIMD machines will be effectively used. However, SIMD machines are not adequate for such intelligent systems like expert systems which should have a sophisticated man-machine interface and programming environment.

4. **Strategies for exploiting multiprocessing capabilities:**
   The current computer usage in molecular biology does not seem to be matured enough to make full use of large scale multiprocessors, although this field has a potential property to produce a great amount of data to be analyzed.

   Using High Level Languages: If these data are once described in conventional languages or informal ways which lack mathematical formalisms, it is difficult to exploit parallelism and apply multiprocessing. Then, the usage of high level languages like functional or logic languages are one of the most important and urgent requirements.

   Providing Easy-to-use Interface and Easy-to-develop Environment: To encourage the usage, the development of easy-to-use and powerful environments has to be started as soon as possible. It may start with small scale multiprocessors and gradually move to large scale multiprocessors having dedicated supports of the advanced database management and knowledge programming functions.

4.4 Networking Capabilities

We computer scientists have learned through our experiences how benifitial to our researches computer network is.

Since the progress in molecular biology is incredibly rapid, computer network must be substantial and important for biologists to come up with and exchange each other

the newest informations, papers and access and share the newest databases and systems finally for bringing out new discoveries and problem solutions.

1. **International Network by 1991 and 1994:**
   Network connection among computer researchers will be extended and will provide better services for the researchers in the world. However, it is not likely that researchers in molecular biology will be able to use computer network systems, for example, a Unix based network, in 1991 as much as the computer researchers use it presently. The key problem is whether biologists will use workstations freely like a paper and pencil. Computer scientists must make more effort to make network systems more friendly.

2. **A central database or local database?**
   Both are necessary. Building of both big central database and local database should be proceed in parallel. If workstations can be distributed widely with special financial supports, local databases will be more important. In this case, one of the key problems will be whether convenient software tools are provided or not on the workstations.

4.5 Database Technology

Besides the software development problem, the database management and access is another important subject to be researched for molecular biology.

The problems on the molecular biological datatbases should divided into two: one is on the internal data representation and structure of the database and the other is on the user interface and tools to access the database.

1. **Internal data representation and structure of DB:** What makes this problem difficult is not only that the amount of data is increasing exponentially every year, but also that there are many unknown corelated factors for each entry.

   Some of the existing databases for molecular biology are based on rather old technology such as relational databases. For such a database like a relational database with a fixed or unique format, the amount of disk storage and main memory required for the process expands in proportion to the product of the number of factors and that of entries. In addtion, join operations on different relational tables occupy most of the computation cost.

   Thus, it is clear that the normal form database bring out the problems of an extravagant amount of storage and unefficient execution.

   Therefore, for both compactness and extensibility of the database and efficient execution of database manipulation, the non first normal form or nested relational databases should be researched and developed to solve or remedy this database explosion problem and offer high performance on database access.

   Among ICOT's database researches, Kappa is a knowledge base management system whose underlying data model is a non first normal form (NFNF) or nested relational model [Yokota88].

2. **Friendly User Interface:**
   There have been several kinds of compact and easy-to-use software proposed for the database management, such as MultiPlan, DbaseII and Lotus 1-2-3 and so

on. Though they are for handling very simple and small databases, but they have been accepted so widely in the business field mainly because of their friendly user interface.

We think of the user interface for molecular biological software as well. Taking a look at the existing software and tools for molecular biology, most of them might work for their own functions but seem to be poor in their user interface. The user interface is a very important factor to decide if the system can be accepted or not. It must be friendly and natural first. The market size for molucular biology is expected to be large enough to research and develop good tools as commertial products.

3. **DB with flexible query capabilities:**
   In the analysis of the correspondence between DNA/RNA/proteins sequences and their 3D structures, advanced database technology with inference functions (knowledge base systems) will be a key for biologists to be able to construct such a flexible databases.

4.6 AI Technology

First, there are many things unknown in this field, only whose phenomenon or facts we can grasp but not whose principle: how they happen. From these unsolved problems, we learn rules little by little through several experiences and expreriments. We might solve the principle at the end, but even if we could, it might be too complicate to represent each rule or process of that.

Thus, the expert system approach in which we take the objective principle as a black box and comprehend it from the visible or well known phenomenon is mostly required in this field.

In other words, the expert systems technology should be applied most effectively and play a great role in the molecular biological research.

1. **Prescreening System:**
   The functional analysis or fuction-to-structure inference of DNA/RNA/proteins may be the most promising field to apply the expert systems technology effectively. It should contain many interesting research topics for computer researchers.

2. **Demon for inconsistency check:**
   AI technology can be used for a variety of analysis in molecular biology. Updating the database with newly found data may raise many interesting research problems in AI. They will include problems in database management, knowledge programming like constraint logic programming, parallel processing and so on.

3. **Relevant AI technology:**
   Pattern recognition technology will also be applicable to analyze raw data on X-ray films and improve the efficiency of biological experiments.

   Natural language understanding may share the problem with the analysis of functions of proteins in reasoning based on ambiguous data.

References

[Aiba88] Z. Aiba, K. Sakai, Y. Sato, D. Hawley and R. Hasegawa: *Constraint Logic Programming Language CAL*, to appear in Proc of FGCS'88, Tokyo 1988

[Chikayama84] Takashi Chikayama: *ESP Reference Manual*, TR 044, ICOT, 1984

[Chikayama88] Takashi Chikayama, Hiroyuki Sato and Toshihiko Miyazaki: *Overview of the Parallel Inference Machine Operating System (PIMOS)*, to appear in Proc of FGCS'88, Tokyo 1988

[Fujii88] Yuichi Fujii, Hirokazu Taki and et al.: *Experimental Knowledge Processing System*, to appear in Proc of FGCS'88, Tokyo 1988

[Goto88] Atsuhiro Goto, Masatoshi Sato, Katsuto Nakajima Kazuo Taki, Akira Matsumoto: *Overview of the Parallel Inference Mahchine Architecture*, to appear in Proc of FGCS'88, Tokyo 1988

[Mukai88] Kuniaki Mukai: *Partially Specified Term in Logic Programming for Linguistic Analysis*, to appear in Proc of FGCS'88, Tokyo 1988

[Nakajima86] K. Nakajima, H. Nakashima, M. Yokota, K. Takai, S. Uchida, H. Nishikawa, A. Yamamoto and M. Mitsui: *Evaluation of PSI Micro-Interpreter*, Proc. of IEEE COMPCON-spring'86, March 1986

[Nakashima87] H. Nakashima and K. Nakajima: *Hardware Architecture of the Sequential Inference Machine PSI-II*, Symp. on IEEE Logic Programming, 1987

[Nakazaki85] R. Nakazaki, A. Konagaya, S. Habata, H. Shimazu, M. Uemura, M. Yamamoto, M. Yokota, and T. Chikayama: *Design of a High-speed Prolog Machine (HPM)*, Proc. of the 12th Annual International Symposium on Computer Architecture, 1985

[Taki86] Kazuo Taki: *The Parallel Software Research and Development Tool: Multi-PSI System*, Proc. of France-Japan Artificial Intelligence and Computer Science Symposium '86, ICOT 1986

[Uchida88A] Shunichi Uchida, Kazuo Taki, Katsuto Nakajima, Atsuhiro Goto and Takashi Chikayama: *Research and Development of the Parallel Inference System in the Intermediate Stage of the FGCS Project*, to appear in Proc of FGCS'88, Tokyo 1988

[Uchida88B] Shunichi Uchida, Tsutomu Yoshioka, Ryoichi Sugimura, Yuichi Tanaka, Koichi Hashida and Kuniaki Mukai: *The Research and Development of Natural Language Processing Systems in the Intermediate Stage of the FGCS Project*, to appear in Proc of FGCS'88, Tokyo 1988

[Ueda85] Kazunori Ueda: *Guaded Horn Clauses*, Concurrent Prolog Vol.1, MIT Press, 1987

[Yoshida88] Kaoru Yoshida and Takashi Chikayama: *A'UM - A Stream-Based Concurrent Object-Oriented Languages -*, to appear in Proc of FGCS'88, Tokyo 1988

[Yokoi84] Toshio Yokoi: *Sequential Inference Machine: SIM - Its Programming and Operating System*, Proc. of FGCS'84, Tokyo 1984

[Yokota88] Kazumasa Yokota, Moto Kawamura, Atsushi Kanaegami: *Overview of the Knowledge Base Management System* to appear in Proc of FGCS'88, Tokyo 1988

A “分子生物学への最新コンピュータ技術の応用”会議の報告

A.1 はじめに

1988年11月3日から5日まで米国シカゴのアルゴンヌ国立研究所(略称 ANL)にて,“分子生物学への最新コンピュータ技術の応用”に関する会議(*Workshop on Advanced Computer Technology and Biological Sequencing*)が開催された.

会議の目的と構成: 本会議の目的は,ANLのMathematics & Computer Science Divisionが米国エネルギー省(DOE)の下で分子生物学研究のために最新のコンピュータ技術を適用することを目指すプロジェクトを発足したいという意図から,分子生物学者と計算機科学者を招待し,プロジェクト発足の準備のための議論と草案の土台作りをすることであった.ANLの要請に対し,国際研究協力の観点から,ICOTから内田俊一と吉田かおるの二名が参加した.

8人の分子生物学者と6人の計算機科学者が招待され,さらにANL内から数人の分子生物学者と数人の計算機科学者が参加した.参加者はその立場もしくは視点から,次の4種類に分類された:

1. 純粋な(計算機とはほとんど無縁な)分子生物学者.
2. 分子生物学をバックグランドとし,計算機システム作りに関与している研究者
3. 計算機科学をバックグランドとし,分子生物学に貢献している研究者
4. 純粋な(分子生物学とはほとんど無縁な)計算機科学者(我々を含む)

Mathematics & Computer Science DivisionのDivision DirectorであるHans Kaperが会議の進行役を務めた.

以下に,会議の概要を報告する.

A.2 会議の概要

A.2.1 初日:チュートリアルと全体会議

チュートリアル: 初日の午前中から3時まで,以下の4種類のチュートリアルが順に行われた.

- **“Sequencing -process and use-” by Prof. C. Woese:** イリノイ大学のProf. Woeseは,ファージとバクテリアに関する分子生物学研究で知られている.

  彼は,分子生物学の基本である,DNAおよびRNAの塩基配列,蛋白質のアミノ酸配列とその構造に関する基礎理論を大まかに述べた.その内容は,第4種(純粋な計算機科学者)を対象にした,ごく基礎的なものだった.

- **“Overview of computational issues from a biologist's perspective” by Dr. G. Olsen:** Dr. Olsenは,Prof. Woeseの下で分子生物学研究のためのソフトウエアシステムを作成している.

  彼は,バクテリア,簡単な真核細胞および複雑な真核細胞を例にとり,現在既知のSequencing Dataの規模と,Sequencing Analysisに要求される計算量およびディスク容量を示した.現在の実験法では,DNA鎖が制限酵素によりフラグメントに分断し抽出されるが,ここで示したデータはフラグメントからDNAの連続鎖に再構成した後からもので,実は,この再構成の過程で組み合わせ論的な膨大な計算量を要するのだと述べた.さらに,分子生物学研究は進化論の解明という“純粋科学発展”の目的のために行われるべきだと唱えた.

- **“State of the art in advanced computing” by Dr. G. Steele:** Guy Steelは,Constraint Programming LanguageおよびCommon Lispの設計者であり,Connection Machineの製造会社Thinking Machine Corp.を先導する一人として知られる.彼は中間層(計算機寄り生物学者)を対象に計算機の最新技術の動向を以下のように分類して述べた.

  - ハードウエアの動向,(Bill Joyの法則:マイクロプロセッサの速度は1984年から年々倍増する$2^{(N-1984)}$)
    - Personal Computer ... Macintosh, Next
    - Engineering Workstations ... Sun, Appolo, Small Vaxes
    - Mainframes ... IBM 370, DEC VAX
    - Supercomputers ... CRAY-1, CRAY-2, ETA-10

* Parallel Computers ... SIMD (CM,DAP), MIMD (Intel, Ametek, Ncube, Topologix)

オペレーティングシステム (IBM, Vax VMS, Macintosh, Apple Domain, Unix, MS/DOS)

- プログラミング言語 (Fortran, C, Lisp, Ada, Pascal) の特質,

グラフィクス標準化 (GKS, PHIGS, Postscript)

- **"State of the art in database technology, expert systems, logic programming" by Dr. N. Goodman:** Dr. Goodman は Codd and Date, International の R&D の Senior Vice President である. 彼は, データベースに関する基礎知識, Sequencing 問題におけるデータベース管理, 情報検索, CAD, グラフィクス, 論理プログラミングの役割を述べた. ヒト DNA の長さは 3×10^9 塩基であり, この生データに加え, 構造に関する情報が加わるものの, データ量およびそれらの検索・更新に要する計算量は, 大量ながら実現可能な大きさであると述べた.

  また彼は, 後の自由討議とグループ討議において, "オブジェクト指向データベース, 高階関係データベースおよび演繹データベースは技術的にまだ未熟であり, また関係データベースは sequence data に不向きであり, 現在の GenBank の実装に賛成する", という極めて保守的な姿勢を示した.

自由討議: チュートリアルの後, 自由討議に入った. 純粋生物学者と我々のような純粋計算機学者が最新技術の導入に関して積極的であったが, 生物学者, 中間層 (計算機寄り生物学者と生物寄り計算機学者), 計算機学者が議論の接点を見つけられないまま, それぞれの立場を主張するのみに留まった. 以下に それぞれ参加者の主張をまとめる.

- **Prof. Casandra Smith** (Columbia University 生物学者, 大腸菌遺伝子の完全な physical map(遺伝子の物理的位置を表す地図) を解明したことで知られる): "皆の目は解析後得られる結果に向けられがちだが, 実験によって得られたジェルの状態の生データは重要であり保存すべきである" と主張した.

  会議修了後, "日本の三島遺伝研究所が握っているヒト DNA に関するデータを世界中が注目している" と語った. また彼女は, この三月に大磯で開催された分子生物学会議のために来日した.

- **Prof. Walter M. Fitch** (University of Southen California, Biological Sciences, 生物学者): 彼はすべての話題に意見を述べていたが, 一貫して何を主張したかったのか判らなかった.

- **Prof. George Church** (Harvard Medical School 生物学者, DNA Sequencing Analysis のためのシステム作り第一人者, ヒト DNA 解明にも深く関わっており, 多量のデータを抱えていると言われている): "Sequencing 問題にとって, ディスクの高速アクセス, メモリの大容量化は必須の問題である" と主張した.

  我々は, 会議出席前の性急な調査でヒト DNA を含む真核細胞の Sequencing Analysis に関して大きな疑問点があった. それは,DNA から mRNA に転写された後 mRNA が核外へ出る際に起こるスプライシング現象に関するものである. mRNA は, アミノ酸にコード化されるコード領域 exon と非コード領域 intron が少しづつ繋がったものであるが, 核脱出時に非コード領域 intron がすべて切り取られ, コード領域 exon だけからの鎖に加工される. これが真核細胞だけ (ただ一つの原核細胞の例外を除き) に起こる現象でスプライシングと呼ばれる. 問題は, どのように exon と intron を見分けるかという点であり, 調べた限りの論文のどれにも述べられていなかったのである. これが判らない限り, DNA から RNA さらに蛋白質の Sequence をどのように機械的に導き出すのか判らないのである. これまでの Sequencing Analysis およびそのためのシステム作りは原核細胞を中心に行われてきたが, 今後の ヒト DNA 解明およびそのためのシステム作りにとって, このスプライシングは重要な点であると思われた.

  会議終了後, Prof. George Church と Prof. Casandra Smith にこの点を尋ねた. 彼らは, "その点こそ彼らが現在抱えている重大な問題である" と答えた. "現在, 三つの方法があり, そのうちエントロピーの変化から認識する方法を取っているが, 現在はまだ完全に機械化されていない" と Prof. George Church は続けた. さらに, "ヒト DNA にとって最も大きな問題のもう一つは, 最終的に得られる蛋白質の構造が大変複雑で, その機能的性質が 三次元構造の表面に位置するアミノ酸であり内部に隠れてしまうものはほとんど影響ない. しかし, そのためには 三次元構造解析が必須であるが, ほとんど手をつけられてないのが現状である", と Prof. Casandra Smith が語った.

- **Prof. Robert Boyer** (University of Texas, Boyer& Moore の Theorem Prover で知られる計算機科学者): "大規模ソフトウエアシステムの開発にとって, プログラミングに比べ, デバッグおよびサブシステムの結合にほとんどの時間を費やされるため, 環境の整備が必須である. プロトタイプシステムと実用システムではその開発時間に 10 倍程どの差違がある." と述べた.

- **Dr. Herve' Gallaire** (欧州における第5世代コンピュータ・プロジェクト ESPRIT の Director) : モデルの構築, 可変構造データベース, 制約プログラミングの重要性を唱えた.

- **Dr. Dennis Benson** (National Library of Medicine, NIH): GenBank を関係フォーマットへ変換する努力がすでになされており, これは使い易い論理設計を行う上で重要である. また, データベースのインデクシング (双方向リンクを含め) はこれからの重要な課題である.

これに対して, 中間層はこれまで彼らが用いた手法 (我々から見たら前時代的なあるいは現在の技術, 例えば, Fortran, C によるプログラミング, 逐次型計算機, Entity Relation データベース などを良しとし, これまで彼らが築いたソフトウエアシステムを守ろうとする姿勢が見られた.

- **Prof. Richard Lipton** (Princeton University 計算機科学者): 特別な技術など要らない. より簡単な技術で押し進めるべきである. 例えば, 関係データベースを導入することにより問題がいかに複雑になるかを考慮すべきである. 並列システムなど Fortran に fork と wait を入れる位でよい.

- **Dr. Tim Hunkapillar** (CalTech 計算機嫌い生物学者, Conway らと大変緊密な研究体制をとっており, 特殊目的のチップを容易に起こせる環境にいる): 特殊目的のハードウエアを使うことは大変意義がある.

我々は, 次のような意見を述べた.

- **内田**: 近い将来, 並列計算機は必須となる. 並列計算機を十分使い熟すには, 高級言語, 特に論理型あるいは関数型言語を使うべきである.

- **吉田**: 議論は, エンドユーザ (純粋生物学屋) の立場に立ったものとシステム開発者の立場に立ったものとに分離すべきである. 新技術の導入は, 言語, マシン, データベースを統合した形で行うべきである.

A.2.2 二日目: グループ会議とグループ別草案作り

二日目は午前中から, 全体を3つのグループに分けてグループ会議となった.

我々のグループは, Nathan Goodman, Dennis Benson のデータベース専門家と Prof. Woese, Dr. Ewing Lusk (ANL), Gail Pieper (ANL Technical Editor) と我々の7人である. 二人のデータベース専門家が, 既存の各種データベースの統合化という路線で議論を進めた. 各種物理データベース上にフォーマット変換ソフトウエアを用意し, 上位ソフトウエア・ツールからある論理レベルとして見える統合データベースを提供しようというものであった. これには何の最新技術も要らなかった.

我々には, 今ならまだデータ量も少ない (例えば GenBank で 50MB) のだから, 将来のためにより効率の良い検索・更新し易いデータベースを作り直すことが可能なのではないかと思われた.

A.2.3 三日目: グループ別草案の併合のための全体会議

最終日, 再び全員顔を合わせ, 各グループの発表となった. グループとしてのレポートを提出できたのは, 我々のグループだけであった. 他の二つのグループは, グループの議論がまとまらず, 個人個人がレポートを提出した.

進行役 Kaper は, グループ共通の目次案を示し, それに合わせて各グループのレポートを再編成するように指示し, 再びグループに分かれてレポート再編成の作業に入った. 三時頃会議は終了した.

A.3 おわりに

3日間の会議を通じて, 生物学者, 従来型技術をベースとする計算機学者, 先端的計算機技術をベースとする計算機学者の3つのグループの間で, かなり激しい議論が行なわれた. 3日間の会議の中では議論は収束せず, 従来型技術支持グループの主張が強く反映されたレポートが作成された.

ANL はその後, このレポートを元に研究所内でも議論を続け, 最終的には並列論理型言語を核言語として分子生物学の研究をサポートする計算機システムを研究開発しようという, 極めて先端的なプロジェクトの提案書を作り上げた [NewScientist]. これは, 5G プロジェクトが後期に展開しようとしている並列推論の技術と多くの共通点を持つものである.

この種の会議は, 日本でも多々行なわれるものであるが, 外国の研究者も招いて立案する開放性, 何の遠慮もなく腹り合うような議論の自由さ, 多くのたような議論の中から本質的なものを素早く抜き出しプロジェクト化する手際の良さなど, 参加した我々は, 議論の中身と共に, アメリカ流の計画立案の方法論を見るという貴重な体験をしたと言える.

今回の会議を"参考"にして提案されたANLの提案書が実施に移された場合，5Gプロジェクトは技術的に共通の研究課題を抱えることとなる．日本の5G技術による世界的貢献という立場から，今後の研究交流の相手として，注目していく必要があろう．

入手資料

ACT&BIO88 *Proceeding of Workshop on Advanced Computer Technologies and Biological Sequencing*, Argonne National Laboratory, November 3-5, 1988

NewScientist New Scientist, November 1988