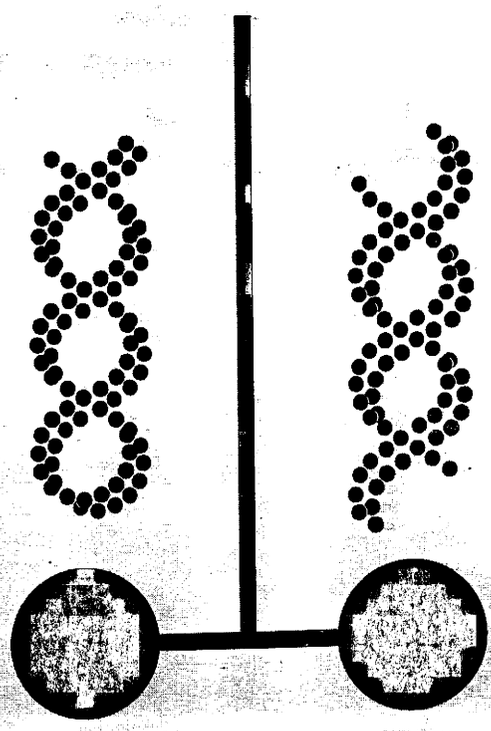
Bio-informatics in Europe

1 Strategy for a European biotechnology information infrastructure



This study has been instigated and managed by The Confederation of European Chemical Industries (CEFIC) in association with a consortium of scientific publishers (Derwent, Elsevier Science Publishers and Springer) with financial support from the CEC.

Extra copies of this report can be obtained from

CEFIC
Bioinformatics in Europe
Avenue Louise 250 bte 71
B-1050 Brussels
Belgium.

Elsevier Science Publishers
Biomedical Division
Bioinformatics in Europe
Room 327
Molenwerf 1
1014 AG Amsterdam
The Netherlands.

Chapter 1

General Goal

Biotechnology, the practical use of biological knowledge, is of enormous importance to the industrialised world and indeed to the whole of mankind. Bioinformatics, the use of computers and information technology in bio-research, is similarly increasingly important with virtually all areas of biotechnology relying, to a greater or lesser degree, on this combination of information and technology.

Bioinformatics is international; results and advances are recorded all over the world and Europe must and does cooperate in this international field. To continue to do so she must have resources that can be committed to international activities as there are signs that international cooperation is breaking down. One reason for this is that the public/private financing structures for biotechnology information and bioinformatics differ between the USA, Japan and Europe. This is leading to confusion and competition where clarity and cooperation are required.

The USA also appears to be developing a national policy in which international participation is questioned. At the same time the USA sells its own, generally subsidised, information services abroad, undermining the local (usually commercial) services and creating a situation where European and other users are becoming increasingly dependent upon the USA for relevant scientific information. This, especially if restrictive practices were later introduced, such as are being mentioned at present, would have far-reaching negative consequences for academic and industrial research and the information industry in Europe.

Europe therefore needs plans to safeguard access and services in this area and to compete and cooperate with the other international players. The best way to do this is to produce relevant data services that can both support the European need and be integrated with similar projects from America and elsewhere to prepare truly international systems.

The goal of this project is thus **The Development of a Strategy for European Biotechnology Information Infrastructure.** This should be both durable and flexible, so it can change in line with the evolution of the field, and guide both the Commission of the European Communities (CEC) and national organisations in decisions relating to the development and support of activities in the area of biotechnology information and bioinformatics.

Chapter 2

Methodologies

This study has been instigated and managed by The Confederation of European Chemical Industries (CEFIC) in association with a consortium of scientific publishers (Derwent, Elsevier Science Publishers and Springer) with financial support from the CEC.

This report of the research phase is based upon a series of interviews with scientists, Research and Development managers and information professionals in industrial and academic institutes in the USA, Canada, France, the UK, FRG, Italy, The Netherlands, Switzerland and Belgium. Additionally, questionnaires were sent to professional end-users and information professionals in all the European Community countries.[1]

Detailed discussions with database producers (including EMBL, GenBank, PIR, HDB, MSDN, MINE, MEDLINE, EMBASE, CABI and BEST)[2] and database hosts (DIMDI, DataStar, SEQNET, and BIKE) have taken place and opinions and evidence taken from library groups and networking organisations such as RARE. Visits to congresses and meetings were made and the relevant recent publications were also examined to follow international considerations such as Japan's attitude to this subject.

Fuller recommendations, with plans for the implementation of a number of policies, will follow in a final report to be produced this summer.

[1] See appendix 1 for the questionnaire and response coverage; a full analysis of the results will be presented in the final report).

[2] A glossary of acronyms is provided in appendix 2

Chapter 3

Scope and Priorities

Our, as yet incomplete, survey of databases relevant to biotechnology shows a wide variety of factual, bibliographic, collections/repositories and directory databases.[1]

The fact that the nucleic acid sequence databases are the foundation for many other types of database, such as protein sequences (so far as these are determined by interpretating the DNA code rather than the direct biochemical protein sequence), and the size of the current international effort on the human and other genome projects, means that most emphasis and interest of people in the field is presently concentrated on the maintenance and development of nucleotide and protein sequence databases and the software required to utilise and interpret this information to the full.

In addition, the need for a comprehensive bibliographic database, linking the various factual and other similar services through a searchable record, is obvious.

Both these represent "hard" or "permanent" data – information that can always be referred to – rather than the more transient data found in directories etc.

The recommendations concerning data banks therefore **concentrate on the sequence and bibliographic databanks** although it is clear that one can extrapolate from these findings to other areas.

Equally important is **easy, secure, and cost efficient access via networks** to databanks in these areas. European users, through Réseaux Associés pour la Recherche Européenne (RARE), are looking at many such specialised needs but the biotechnology community is already developing dedicated services to facilitate the use of the above types of data; these have also been examined.

[1] this list will be published with the final report.

Chapter 4

The situation in the USA and Japan in relation to Europe

The USA has made a firm commitment to biotechnology information and has, in the recent past, developed far-reaching bioinformatics policies and funded these in a coordinated, long term, manner. Many US agencies and government departments are involved but the commitment to this policy is perhaps best characterised by the establishment of a dedicated centre to this subject – the National Center for Biotechnology Information (NCBI), which is closely allied to the National Library of Medicine (NLM) of the National Institutes of Health (NIH).

The US, and other national governments, are looking closely at the resources needed for the mapping and sequencing of the human and other genomes. This effort will result in a series of international sequence and related non-commercial databases.

American public funding in this region is large. The NCBI alone had a budget of $8 million for 1989 and the NIH had reserved $27.6 million for Human Genome work. The Bush administration has asked for $128 million for the entire Human Genome project in 1990 but other estimates indicate that total US Federal spending in this area will be $200 million a year leading to a total budget of around $3 billion. Estimates are that the final budget for information services within the Human Genome project will be in the order of $30 million per year at the height of that project.

This programme will give other areas of biotechnology a great impulse as the genome activities will stimulate the development and use of parallel programmes such as other nucleic acid sequence and genetic mapping projects and protein structure and function activities.

There are also a number of Human Genome projects in Europe which will stimulate bioinformatics as well as biological research . The UK has just begun a major initiative where informatics and education will play a central role. The European Community will also have its Human Genome Analysis Programme. This, too, while not concentrating on sequencing per sé, will provide a great deal of data that will find its way into specific and general databases. Firm arrangements for merging this material with other international programmes are required.

The present nucleic acid sequence databases (normally referred to as GenBank/EMBL) are built from international inputs derived, presently, from Europe's EMBL (45%),

Japan's DDBJ (5%), and the US Los Alamos team (50%). IPIR, the international cooperative of the PIR project, is similarly built up by the Max Planck Institute for Protein Sequences at Martinsried (MIPS) supplying 35% of PIR data, the balance being American 50% and Japanese 15%.

This international coverage is achieved in a variety of ways. EMBL and Los Alamos have divided the source literature between them, they exchange the collected data between themselves but produce two separate databases in slightly different formats. PIR is one database and receives input from various European, Japanese and American sources.

As part of the streamlining taking place in bioinformatics in the USA the NCBI is taking steps to take over the production of the American sequence database GenBank from Los Alamos National Laboratory, and may well play a far greater role in the funding and production of the American side of the Protein Identification Resource (PIR) currently produced by the National Biomedical Research Foundation (NBRF). While this might well aid the American effort it is causing confusion in Europe.

The NCBI plans envisage using MEDLINE to identify those articles with relevant information for the specific databases. In principle this means the NCBI will cover **all** the literature (although there are many indications that using MEDLINE alone will not be sufficient to locate all the required primary articles). These moves could have far-reaching consequences for the international nature of these projects; and others. If the Americans are covering "the literature", in fact the relatively easy to locate, central core literature, then the other centres will be left with collecting the more difficult to locate and therefore more expensive peripheral information if complete international projects are to be built. No real international thought appears to have been given to this difficulty and so no clear lines for future collaboration been properly debated. Furthermore, there is no clear-cut authority in Europe to negotiate with the NCBI on these matters.

International science recognised the need for comprehensive and mutually supportive databases some years ago and CODATA, the Committee on Data for Science and Technology of the International Council for Scientific Unions, planned the Hybridoma Data Bank (HDB) as a collaborative venture with Europe, the USA, and Japan each producing relevant data from their region and exchanging this via a central accreditation centre. The nodes then receive a collated tape back for local use and exploitation and, in this way, produce a database without a dominant owner or exploitant.

Each node is financed by national or regional grants but the European node, in 1989 supported by CEC and national moneys, has to recoup as much expenditure as possible through direct sales of the data. However, as neither of the other partners is in the same position and both can sell, or even distribute free, via electronic means, the same file into Europe, the future of the European node is threatened; there is little evidence that European users, while finding the idea of American dominance undesirable, are actually willing to pay more for the **same** information or service.

We therefore have a "financing conundrum" where the funding partners apply different rules to the participants and where it is increasingly difficult for European partners to gain revenues from their exploitation of any joint file. This difference in funding methodology could have far-reaching consequences for Europe's role in international database activities.

Thus neither of these models offers, at the moment, long-term structures for the preparation of international databases.

Bibliographic databases have tradition-

ally been used to locate the relevant primary article. Europe has many biotechnology-relevant commercial services, as does the USA, but the Americans also have the MEDLINE series of databases produced by the NLM with grants from the NIH. This subsidised product has long threatened the commercially produced, non-subsidised, European files but, increasingly, such bibliographic databases are gaining another role as they will be used to identify the relevant primary articles for uptake into such specialised data collections as the nucleic acid sequence databases or PIR or the HDB.

This is part of the basis for the NCBI taking over the American sequence databank building activities: they will identify the relevant papers by scanning MEDLINE. MEDLINE will then probably be further integrated/cross-referred to a number of other databases in the future, to serve as the pivotal point for US biotechnology and biomedical data services.

This will further undermine European bioinformatics. Due to the NIH subsidies, MEDLINE has become the first choice bibliographic biomedical database world-wide; but it does not always satisfy European needs and increasingly cannot cover all the relevant and important European primary information sources. These moves further threaten European bibliographic services by highlighting MEDLINE's central gateway role and mean that European research results, published in non-MEDLINE-abstracted journals, will be lost from the MEDLINE-derived databanks. This, in turn, could lead to the better research papers migrating from those European primary journals not abstracted in MEDLINE to their American counterparts. It will also damage international bioinformatics, as the completeness of the various activities will be threatened.

All these services increasingly use **networks**. The NLM is actively developing its own telecommunications network so that users can reach various databases. BIONET, a non-profit resource run for the scientific community, funded through the Research Resources Division of the NIH, was, until recently, the main network for sequence-related work and offered international access to international destinations such as GENBANK, EMBL, PIR, SWISS-PROT etc., as well as offering mail and Bulletin Board services. The funding for this is being stopped and the service will, presumably, be replaced by a presently, American, NIH/NLM alternative.

A large majority of European scientists, from all parts of R&D, want a common telecommunications network. Most academics presently use their own national academic network and these are interlinked through EARN in Europe and BITNET in the USA. Technical improvements are required but progress in this direction is felt to be slow and, in the absence of a common service, we are seeing the emergence of a number of subject-defined networks.

One example concerns the European sequence area which has just begun to be served by EMBnet, a system linking the EMBL Data Library with a number of national/regional and industrial users. This service allows the peripheral nodes to download a nightly update of new nucleic acid sequence data. These peripheral nodes are increasingly carrying the required software to handle the EMBL and other related files and EMBnet is developing into an integrated service with two-way communication between the collating centre and peripherals with their specific needs. It is able to interact with many other similar services.

Some peripheral nodes reached via this and other networks, are serving a national function. The Centre Inter Universitaire de Traitement de l'Information Banques de Données Biomedicales (CITI-2) at the University of Paris offers users current awareness services and the opportunity of entering sequences for

local checking before submission to EMBL. SEQNET (Sequence Network Daresbury Laboratory UK) offers similar services and access to a number of databanks (e.g. EMBL, GENBANK, PIR, SWISS-PROT, Brookhaven). The European Human Genome Analysis Programme is planning a similar service to EMBnet and it is clearly advisable that these similar services be integrated and mutually supportive.

Other, non sequence-related, international communication networks are developing to improve communication and service among biologists. CODATA has established an electronic mail service on the commercial Dialcom service which offers users access to a variety of culture collection directories and bulletin boards; the Microbial Strain Data Network (MSDN) is also available on Telecom Gold (the UK owner of Dialcom) and offers access to a variety of culture collections and directories.

In conclusion Europe is in no way lagging behind the USA in the intellectual or basic technical expertise required in individual areas. However, bioinformatics in the USA is more centralised with at least the NCBI offering a focal point for research and development and the subject enjoys a greater political commitment. It has a firm financial basis through continuous funding while, in contrast, Europe has no central policy and lacks long-term structures on which funding can be based.[1]

This lack of continuity certainly restricts the development of new data services and European centres of excellence. The difficulty of finding money and the lack of even medium-term security makes it difficult to develop new services against an already highly uncertain and dynamic background. This is undermining the moral and long-term commitment of the scientists involved and is threatening Europe's present position of near parity in terms of data input, thereby weakening Europe's influence on international biotechnology information activities. These weaknesses will lead, in turn, to an increasing dependence on American data services and a further chain of events ending in renewed calls in Senate, Congress and among industrial and even certain academic groups to restrict access to biotechnology information to American users only.

Certainly most interviewed scientists feel there is room for a lot of improvement in Europe and there is a general realisation among the leading researchers that America is purchasing its way into a potentially dominant position by better funding and by "buying away" key European staff. Few long-term bioinformatics recruits from America to Europe can be found but many leading researchers have crossed the Atlantic from Europe to develop bioinformatics projects. This trend, and the precarious base on which European bioinformatics is currently based, is not generally recognised by European users who are happy to receive American subsidised material and even wonder why the (non-supported) European material is so expensive. However, when the potential dangers of restricted or delayed access are explained, European academic and industrial users regard the situation as being highly undesirable.

[1] Examples include the fact that the European node of the international Hybridoma Databank (HDB) requires 150,000 ecu from public money in 1990 to continue its role in this project but there is no "open budget" to provide this and the team are being forced to look for research-based grants to continue what is really an infrastructure task. The EMBL Data Library has recently also had to be temporarily "rescued" by last minute injections of additional money from the CEC above the continuous support from the European Molecular Biology Laboratory. The lack of infrastructure money means that this vital service is being funded on research criteria.

Chapter 5

European User Needs

All interested parties agree that the biotechnology-relevant databases should be open to the international research world.

Similarly, all scientific users of basic nucleic acid sequence and related data feel that the basic scientific data is essential as a **research tool** and that this must be available without restriction of access and therefore funded from public money. This is vehemently the case among academics who are very concerned lest European services become wholly commercial although scientists in industry are more willing to see such services funded from private sources.

There is a complete rejection of any system that will require scientists to pay for raw, unmanipulated, data while they themselves are submitting such data to the same database(s) without remuneration. The intensity of use of these databases is, moreover, directly proportional to completeness and timeliness; it is therefore essential that an infrastructure be supplied that allows easy and cheap access to the basic, raw, material.

This raw data can be handled only with specialised software. A majority of correspondents feel that an European software clearing house is required.

Users increasingly require more than one source of information. Services such as Integrated Sequences and Structure (ISIS – a protein structure database product produced in the UK), analyse data from several basic databases – in the case of ISIS, seven. Such projects are highly specialised and require development investment which, on most occasions, cannot be recouped according to normal commercial rules. There is therefore the need for a combination of public and private funding in this area – at least while it is evolving.

There are many specialised data services derived from the original published article. This implies a need for **basic data** that can be used by specific centres and/or services to provide **added value products**. Some value is added merely by combining different types of data together in a suitable format and most if not all factual data must be cross-referenced back to other forms of information (such as the primary article, or the relevant microbial strain or cell line). There is an acceptance that such "added value" can be charged for .

Europe would benefit from establishing a series of inter-connected information centres with certain, basic or core, databases maintaining a global coverage of the specific data in its most simple form. Other databases/centres would take this data, adding value in the form of better annotation or other improvements. The "basic services" will need to collect and

correlate data from all international sources and so must ensure that they work in association with other similar services. (An example could be a central depository of nucleic acid sequence data with basic annotation where other more specialised peripheral centres might "add value" by adding *specific* annotation.)

It seems likely that the need for the more specialised services will be handled within Europe on subject specific or national/regional data services (e.g. a databank of human genome sequences for all Europe, or a national centre satisfying pre-market-researched database and service needs such as SEQNET at Daresbury in the UK). It is also essential that such European services are integrated with other international services; ultimately scientific information needs to be as exhaustive as possible and the best way to achieve this is through international collaboration.

Training and "HELP" services are few and difficult to find. Very few academics have been formally trained in bioinformatics and only the larger industrial companies appear to be able to concentrate on the proper training of staff. One or more European centres specialising in a particular aspect of bioinformatics could be considered (this could be combined with one central service or with a software centre). National language services, or at least Help and Advice Centres, are also needed but there is a totally overwhelming insistence that English be used as the bioinformatics language. Any national service should be linked to regional or national nodes of networks established to service either local markets or to regional scientific or medical libraries if they exist. Such services are especially needed by smaller academic and industrial centres who lack the specialised staff and resources needed to use the increasingly sophisticated information technology products.

Industry requires the same data and services as academia. Certain companies also require a high degree of security with regard to what they are searching for and on. Thus Europe must offer adequate stand-alone as well as online services; and must allow access to the various international sources on commercial as well as academic networks.

Chapter 6

Conclusions and Initial Recommendations – Towards a European Bioinformatics Strategy

There is no clearcut European policy in bioinformatics although Europe has a number of national and international activities which, **together**, produce an impressive total. A number of European national government initiatives are under way, especially in the FRG and UK. BICEPS, funded by the European Commission through DGs XII and XIII, contributed to bioinformatics elements in the five year "Biotechnology Action Programme" (BAP) which are being subsequently further developed in BRIDGE; and to the launching of the Advanced Informatics in Medicine (AIM) programme.

Europe also has many excellent and relevant commercial products that, with the correct stimulus, could be used alongside publicly funded files for the further good of the total market.

All these projects and programmes can and should be better coordinated and should be brought under one strategy so that they support each other.

This is especially the case where databases were/are developed as an integral part of a research project by specialist teams who combine the scientific and computing expertise required. These are therefore often institute-bound but bioinformatics is too important for it to be subject to the uncertainties of short-term research grants or the goodwill of individual institute policies. Essential bioinformation services should be placed in institution-independent structures and controlled and financed by the CEC in association with other relevant European and national funding agencies.

A coordinated framework, within which the present and future stand-alone projects can be managed and disseminated, is therefore required. Such an European strategy should be based on a willingness to establish standards and protocols for collaboration on an international scale.

In this regard two passages from the June 1989 Report of the Framework Programme

Review Board[1] seem worthy of quotation.

> Fragmentation of the European research environment is paralleled by fragmentation of research policies and ensuing lack of coherence in long range strategic objectives for research, especially in those areas that can benefit from large scale cooperation in planning.
>
> Existing European cooperation in science and technology adopting variable geometry (under, for example, EUREKA, EMBO, ... ESA, ESF-networks, etc.) should not be replaced, absorbed or duplicated by Commission sponsored programmes. Synergistic relationships should be built up between such programmes and the community R&D effort on a case by case consideration of the most suitable arrangements and need.

Bioinformatics provides such a focal area as it is central to a large number of European research activities in the strategically important life sciences and technology market.

The CEC, seen against the total of all the European national programmes, might not play a dominant role in this area but it can and should play the essential position of coordinator and stimulator, providing the continued technical, legal and financial infrastructure as well as specialised services required. By **infrastructure** we mean the technical (e.g. networks) and basic scientific information services (such as a central bibliographic file) upon which other, sectoral, data services can be built (i.e. areas of specific scientific interest). In doing this the Commission should ensure that Europe as a whole benefits from the constituent efforts in this area. Both DG XII and DG XIII should be involved but should complement rather than compete with each other.

The CEC should further use its influence to ensure that scientific information remains freely accessible and, wherever possible, as part of a truly international service.

Bibliographic Databases

The bibliographic database remains the key source to locate the original article and/or information source. The Americans, and some European services, rely on MEDLINE but this does not satisfy all the European needs (the coverage of European journals is not complete so some European work is "lost" to the secondary services).

There is also an increasing tendency and need for authors to directly deposit sequence and other data into data banks. This data must not be lost from the present services. Steps should be taken to ensure that Europe's relevant services are attractive and competitive so that customers are motivated to use them.

The European commercial, and relevant non-commercial database producers should be encouraged to collaborate. This could be to produce a **common core database** of citations that can be used by all database producers to reduce duplication and cut costs thereby improving competitiveness while providing a core resource for academic, commercial, and non-commercial services to produce added value databases.

The CEC should not seek to establish another central bibliographic database to support other specialist data services that would compete with existing systems.

European national governments and the CEC should ensure that fair and legal conditions are developed for the operation of all bibliographic and related databases in the in-

[1] prepared for Vice-President Pandolfi, The Commission of the European Communities, by Pierre Aigrain, Sir Geoffrey Allen, Eduardo De Arantes E Oliveira, Umberto Colombo and Hubert Markl

ternational market place.

Factual databases and collections

Many bioinformatic products originated in Europe (e.g. EMBL was first to launch a nucleic acid sequence databank). However, while start-up money can sometimes be obtained in Europe, usually as part of a research grant, continuing funding is difficult. Bioinformatics has to be seen in the light of **cost/benefit** rather than **profit/loss** economics. The loss of such information will mean a loss of competitiveness in academic and industrial biotechnology; the benefits are generally widespread, diffuse, long-term and difficult to measure in identifiable commercial terms but are nevertheless great, continuing and of cummulative strategic importance.

The commercial organisations cannot be expected to compete against subsidised American competition, and neither the academic nor the industrial market is willing to pay more for a project they can obtain almost "free" from the USA. Therefore European governments and/or the CEC should define responsibility for providing these core services in nucleic acid sequence and other basic, primary, databases. Given the present different granting structures between Europe and America and Japan this will require new funding rules as well as firm agreements with any future partners on the exploitation of publicly funded material. The CEC should ensure that European opinion on such questions as "the free access to scientific information" and "the exchange of relevant data" is coordinated and well presented in international negotiations.

In principle Europe only requires one central collection and collation centre for basic nucleic acid and another for protein sequence information (although these tasks could be carried out in one centre). Current policies allow, and even promote, the duplication of similar activities in these similar centres with resulting inefficiencies of duplication. The Nucleic Acid Sequence Data Library, currently situated at EMBL, has a long history of innovative work. Their product and services are respected and regarded as essential and this expertise should therefore be maintained. MIPS is developing expertise in proteins and could become the European centre for this area. The two centres should be supported but managed in such a way that they complement each other; there are insufficient funds to allow them to compete.

These core databases will supply data for other services but should not be seen as the only centres where databases in genetics and molecular biology can be produced. Thus specialised activities, such as annotating the Human Genome, can be delegated to other centres so long as they are in direct contact with the Central Data Library and ensure that Europe maintains **one** complete basic data set. The expected expansion of data means that the present facilities at EMBL might become too cramped within the near future. Europe might then require a purpose built facility for this service. One could then examine whether this centre might play a coordinating, or even a physical, role in maintaining the various other bioinformatics services such as the Hybridoma Databank, Carbank, MINE etc.

Other specialist data services must also be readily available in Europe. A number of databases currently only available in America such as the Genome Database (GDB), Baltimore, OMIM (Online Mendelian Inheritance in Man), Baltimore, and the Mouse Genome database in Jackson are further examples of the range of material that is required for efficient research. These should be readily available in Europe.

Certain specialised products and activities in the bioinformatics market (e.g. annotated sequence databanks, ISIS) are still too dilute (the data is thinly dispersed across the field) and immature for commercial exploita-

tion. Yet these databanks, software applications and directories are essential for biologically based R&D. This report shows that measures are needed to guarantee the financial stability of basic services. There is also a need for initial financial support for services that might later become self-supporting, even if these are not research based but as long as they support the research need. The CEC should ensure that the present funding structures are changed so that infrastructure projects are continuously funded.

Networks, Hosts and Software

Sequence data bases require an efficient communication network to receive and distribute relevant data to national or regional and industrial nodes. EMBnet, linking as it does the EMBL Data Library to a number of national and industrial nodes, satisfies these needs in the sequence field and should be supported. This could be done by the CEC supporting the central node, leaving the peripheral nodes to be financed by special interest, commercial, or national funds (e.g. SEQNET in the UK). EMBnet should also be interlinked with other similar services so that Europe is linked by an integrated set of networks dealing with specific areas.

EMBnet should not per sé become the **dominant** network. Other specific services, such as the MSDN, are developing and the CEC should encourage these various services to interact and interlink so that industrial and academic users can ultimately communicate with each other on one European bioinformatics network or via interlinking gateways. The continuity of such networks is important and this topic will be covered in more detail in the second report.

The commercial database hosts have little interest in offering bioinformatics data at present. National nodes, such as SEQNET, could offer an increasing number of relevant files and services especially in conjunction with the commercial database producers who should be encouraged to produce relevant files to support the factual databases.

Bioinformatics is dependent upon good software. Europe would benefit from a central software clearing house that could advise and train academic and industrial users.

Appendix 1

The questionnaire was distributed by mail through CEFIC and various national organisations. In some countries meetings with interested parties were organised before the questionnaires were distributed. All concerned were encouraged to distribute more copies of the questionnaire so that the final exact number of despatched copies is unkown; it will be more than 300 but probably less than 400.

135 answers had been received by the end of february 1990, thus a coverage of about 35%.

The responses were not evenly spread across the various countries. Belgium has been very active and accounts for some 35% of the total returned. France and Denmark returned fewer than expected but the FRG, Italy, Netherlands, UK and Switzerland all returned between 10 and 20 forms from which an acceptable insight into the national situation and user needs could be assembled.

The response from industrial companies (more than 40 answers) is relatively high compared with the coverage of the university and institutional centres in Europe.

A full analysis of the questionnaires will be given in the final report.

Industry respondents

Belgium

Amycor, Clovis Matton, Cobrew, International Bio Synthetics, Labofina, Plant Genetics Systems, Smith Kline Biologicals, Solvay, UCB.

FRG

Bayer, BASF, Behrily Werke, Boehringer Mannheim, Hoechst Biologische Forschung, Hoechst Pharma Forschung, Hoechst Zentral Forschung.

Denmark

NOVO-Nordisk.

France

Rhone Poulenc Santé, CEN Saclay, Sanofic Elf.

Italy

Eniricerche, Farmitalia Carbo Erba, Merrel Dow Research Institute.

The Netherlands

AKZO Pharma, Biores/Anglian Biotech, Dalton, DSM, Duphar, Gist Brocades, HBT Holland Biotech, Zaadunie.

Switzerland

Ciba-Geigy, Hoffmann La Roche, Nestlé, Sandoz.

The UK

ICI Agro, ICI Pharmaceuticals, ICI Seeds, Beecham Pharmaceuticals, Glaxo, Nickerson International Seed Company.

Appendix 2

AIM Advanced Informatics in Medicine, CEC programme.

BAP Biotechnology Action Programme of DG XII.

BEST British Expertise in Science and Technology – an expertise database listing researchers and their skills.

BICEPS Bioinformatics: Collaborative Programmes and European Strategy.

BIKE Biotechnological Information Knot for Europe – a (German) information service maintained at GBF Braunschweig.

BITNET the American research community telecommunications network, effectively the same as EARN.

BRIDGE Biotechnology Research for Innovation, Development and Growth in Europe. DG XII programme.

CABI Commonwealth Agriculture Bureau Information database – a large database covering agriculture and agricultural biotechnology.

CARBANK A international carbohydrate structures database, presently in preparation.

CEC Commission of the European Communities.

CEFIC European Council of the Association of Chemical Industries.

CERDIC European Centre for Research and Diffusion of Immunoclones.

CODATA Committee on Data for Science and Technology of the ICSU.

DDBJ DNA Database of Japan, National Institute of Genetics, 1111 Yata Mishima, Shizuoka 411, Japan.

DIMDI Deutsches Institut für Medizinische Dokumentation und Information.

EARN European Academic Research Network, a telecommunications network linking European academic networks with each other and with American services.

EMBASE Online version of the Excerpta Medica bibliographic database, produced by Excerpta Medica, a division of Elsevier Science Publishers, Amsterdam.

EMBL European Molecular Laboratory, Heidelberg; funded by fifteen European member states. EMBL produces the European Nucleic Acid Sequence Databank.

EMBO European Molecular Biology Organisation.

EMBnet European Molecular Biology Data network – network connecting many university and other computers, primarily for the dispersion of the EMBL data library.

ESA European Space Agency.

ESF European Science Federation.

GDB the Genome Database available in Baltimore, USA.

GENINFOA nucleic acid sequence database with pointers to other relevant databases produced by the NCBI.

HDB The Hybridoma DataBank, an international databank covering hybridomas and other related subjects.

ICSU International Council for Scientific Unions.

ISIS Integrated (protein) Sequences and Structures, produced by Leeds University, UK.

JIPID Japanese International Protein Information Database, Tokyo, Japan.

MEDLINE The online database of MEDLARS produced by the National Library of Medicine, Bethesda, USA.

MINE Microbial Information Network in Europe. A network linking a series of culture and related collections across Europe.

MIPS Max Planck Institute for Protein Sequence Data at Martinsried. The institute produces the European input for the international Protein Identification Resource database (PIR).

MSDN Microbial Strain Data Network. An electronic mail service connecting users to various culture collection and related databases. Runs on Telecom Gold, a commercial electronic mail service.

NBRF National Biomedical Research Foundation, Georgetown, Washington, USA. Producers of the (International) Protein Identification Resource database IPIR.

NCBI National Centre for Biotechnology Information, Bethesda, USA.

NIH National Institutes of Health, Bethesda, USA.

NLM National Library of Medicine of the NIH.

OMIM Online Mendelian Inheritance in Man database, Baltimore, USA.

PIR Protein Identification Resource (of the NBRF).

RARE Réseaux Associés pour la Recherche Européene.A CEC supported team researching the various telecommunication needs of Europea's research community.

SEQNET (Nucleic acid based) series of databases carried on the SERC Daresbury computer (UK) and made available with supporting software to academic and industrial customers. A node on the EMBnet service.

SWISS-PROT Protein sequence database by Amos Bairoch, University of Geneva, in collaboration with the EMBL data library.