HOW LONG IS CO-OPERATION IN GENOMICS SUSTAINABLE ?

P.B. Joly; V. Mangematin[+1]
INRA/SERD
Université Pierre Mendes France, BP 47X
38040 Grenoble
Tel : (33) 76 82 56 86
Fax : (33) 76 82 54 55
E-Mail : vincent@grenoble.inra.fr

Publications on the 16 yeast chromosome sequences group together over 400 different authors from Europe, Japan, Australia and the USA. When research is not organised in networks, it is carried out in large sequencing centres such as the Sanger Centre in Britain, the Helix Institute in Japan or Saint Louis University in the USA. Both cases illustrate the collective nature of knowledge creation. Other examples of co-operation between numerous researchers in various countries, more closely related to innovation, might also be mentioned, such as the development of software for comparing proteins or DNA sequences.

Collective publications reveal the collective nature of research, whether it is carried out by major consortia (the case of yeast) or around large research facilities (such as the synchrotron or major genome sequencing centres). This collective nature stems from two factors: (1) the advantages of co-ordinating efforts on major projects (e.g. economies of scale and of collection) and (2) very strong interdependency in the creation and utilisation of knowledge (related to cumulativeness).

The detailed analysis of scientific practices in genome research (yeast) during the years 1987 to 96 provides suitable material for identifying the characteristics of scientific production in emerging fields, i.e.: (1) the collective nature of scientific production,

[+] corresponding author
[1] We would like to thank A. Goffeau and P. Mordant for their helpful coments. Of course, all errors remain our own.

revealed by the number of authors of publications; (2) the co-ordination of researchers around a handful of major facilities (sequencing centres); (3) partial or delayed publication in journals or proceedings; (4) numerous controversies on obligations to divulge information[2] or on the referee system[3]. From lawsuits to controversies, and from controversies to anecdotes, genome research offers the image of a world in which tension between the creation and the utilisation of knowledge is strong.

The first section of this text analyses the principles underlying the organizational system in Europe. It describes the conditions under which this type of approach is effective. It insists especially on the financial conditions which are up to 4 USD per base. Some elements of impact of this organization on the relative position of science in Europe are presented. Although the European experience of yeast sequencing is an interesting one, it cannot be imitated exactly. Hence, the second section examines the extent to which certain modes of organization warrant being adapted and transposed to the sequencing of other organisms.

1. EUROPEAN ORGANIZATION OF YEAST SEQUENCING

The entire scientific community celebrates the complete sequencing of the Saccharomyces cerevisiae genome and the making available of genetic information and material. Indeed, yeast is an important organism for biologists because it is used as a model (Vassarotti, A., Dujon, B. *et al.*, 1995). The E.U. financed 56% of the sequencing, which was carried out by almost a hundred laboratories specialised in biology and therefore having limited sequencing capacities.

The European system is in fact radically different from that in America, Canada or Japan, where sequencing capacities have been concentrated in a small number of laboratories. The mode of organization chosen by the E.U. offers original answers to recurrent questions on the conditions of appropriating results, the setting up of financial and scientific incentives, the division of work and the public or private nature of the knowledge and artefacts produced. A detailed study of the division of work between European laboratories has revealed the main principles underpinning the co-ordination of activities. It shows how the

[2] Scientific journals compelled researchers to submit oligonucleotide sequences to one of the public banks before publication. This practice was jointly decided by about fifty journals at the beginning of the 90s. Nature joined them in January 1996.

[3] The establishment of its rules of good conduct followed the fraudulent use by one of the reviewers of the prestigious journal Nature, of data in an article that had, moreover, been refused publication (Science, 270, 22 december 1995, p 1912).

various incentive mechanisms and legal and contractual aspects combine to ensure that all the actors profit from the research effort.

European organizational system

In 1988 the various teams involved agreed on a global, multi-national strategy, i.e. repartition of sequencing by chromosome; step-by-step sequencing; separation of functionalization and sequencing; establishment of quality standards; and work on an identical strain provided by M. Olson and L. Riles (Saint Louis, USA). The standardisation of results enabled the teams involved in the sequencing to organise themselves as they wished, and to use diverse techniques.

The organizational system adopted in the EEC, based on these guidelines, was coupled with a number of rules which have promoted co-operation between over a hundred laboratories (Dujon, B.e.a., 1994; Feldmann, H.e.a., 1994; Oliver, S.e.a., 1992). As the diagram below shows, the co-ordination of work by chromosome is entrusted to a researcher who is responsible for sequencing on the entire chromosome and allocates tasks to the various co-contractants. He/she works in collaboration with the informatics co-ordinator who helps him/her in allocating the cosmids at the start and assembling the sequences at the end.

Diagram 1 : Functioning of the system*

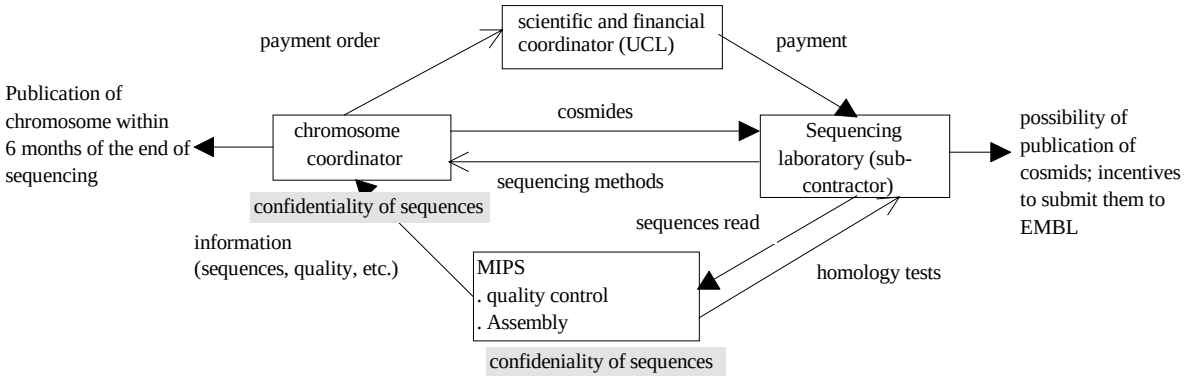


Diagram 1 shows how the different elements in the organizational system interact: financial incentives, scientific incentives, guarantee of confidentiality and dissemination of genetic material and information.

* MIPS : Martins ride Institute for Proteins Sequences. Informatic coordinator. UCL : Université Catholique de Louvain

Principle 1 : Standardisation of results

The standardisation of results is a core element in the coordination of the division of work in Europe. It facilitates the establishment of financial incentives based on quantifiable results. The terms of reference define very precisely the respective role of each of the chromosome co-ordinators, information co-ordinators or laboratory sequencers. They define mainly the type and status of the objects exchanged, e.g. quality standards (ratio of errors in the sequencing of cosmids); minimum sizes of cosmids - which increase progressively to facilitate the management of sequencing (from a set of 8 Kb in 1989 to 2 or 3 sets 35 Kb in 1995); and formats for computer data. On the other hand, the laboratories are entirely free to choose their sequencing methods.

Principle 2 : Payment by the piece and rate differentiated in relation to quantities

Financial incentives exist on several levels: sequencing laboratories, chromosome co-ordinators and informatics co-ordinator. In each case the principles are similar: payment by the piece and tapering rates in relation to learning. Payment by the piece is one of the ways in which laboratories are forced to honour their commitments. It is possible only when expected results are defined precisely and are measurable. In our study set rates varied according to technical progress and to experience gained. Hence, in the first biotech (BAP) contract in 1989, the price paid to the sequencing laboratories was 5 ECU per pair of bases, while in 1996 it is 1,6 ECU.

Rates are differentiated in relation to the quantities sequenced. The price per pair of bases paid by the EEC to a laboratory is lower if the laboratory exceeds a certain volume of contiguous sequenced cosmids (e.g. 100 Kb in Biotech II). This reduction of the unitary rate constitutes a ceiling on the laboratories' profits. Thus, costs were calculated to limit situation rents without discouraging potential entries in sequencing (rate differentiated in terms of quantities sequenced). Whether an outcome of financial incentives or the discovery of the necessity of acquiring know-how in sequencing, the number of laboratories involved in the sequencing of yeast has risen from 35 (chromosome III) to over a ninety (total number of chromosomes sequenced by the EEC). Some of these joined the Biotech II programme with very small quantities to sequence (25 Kb) [cf. biotech contract]. Tapering rates thus make it possible to avoid the establishment of barriers to technology entry, and to limit the effects of learning.

Principle 3 : limited priority rights and incentives to publish

In collaboration with the MIPS, the co-ordinator is responsible for publishing the chromosome sequence when it is complete and when quality control, assembly and verification have been carried out. Such publication in prestigious journals is co-signed by all the participants, with the chromosome co-ordinator being featured as the main author and the MIPS manager as the last author.

Publishing by chromosome is not systematic. The Americans and UK, for example, made public information by groups of cosmid (100-200 kb), which has enabled them to disseminate information faster. The European procedure explains delays of up to two years in publishing information. However, according to the stakeholders, these delays enable them to guarantee the quality of the information (absence of sequencing errors) [Goffeau, personal interview, 6/96]. Submitting the sequences to the MIPS and publishing data are clearly two separate procedures. Laboratories are strongly encouraged in the following ways to submit their sequences to the MIPS: the "first come first served" rule is an incentive to research teams to submit their sequences rapidly in order to receive others; in the case of overlapping parts priority is given to the first team that submits its results; a part of the payment (50%) is subject to conformity with quality requirements. By contrast, incentives to publish sequences are merely ethical since the sequences submitted to the MIPS are covered by a confidentiality clause until publication of the chromosome.

Between submittal to the MIPS database and publication of the entire chromosome, the laboratory can exercise *a priori*ty right in so far as it has complete latitude for carrying out complementary biological research and even publishing its results or patenting them. The allocation of a segment to a laboratory is therefore attended by temporary "ownership" (or a right of reservation) which theoretically ought to act as an incentive to link sequencing activities to complementary activities oriented towards application:

> "Each segment assigned to and accepted by a participant becomes his/her property for the duration of the sequencing of the entire chromosome and cannot be claimed by others" (Vassarotti *et al.*, 1995: 133).

Yet, any sequencing laboratory can ask the Mips to compare a sequence to the sequences of the base yeast even if the latter are confidential. If homologies with unpublished sequences are identified, the two laboratories are put into contact.

> "A database of the confidential, unpublished but submitted sequences is indirectly accessible by the user community: requests to search for homologies in the confidential database are processed individually and matches to confidential data are indicated. Both the laboratory that submitted the

> sequence and the scientist who issued the request are informed of the existence of similar sequence" (Vassarotti *et al.*, 1995: 134).

The contractual system is new and interesting. The main concern here is to give participant laboratories the opportunity of exercising a dual function: that of sequencing, which is codified precisely, and that of complementary biological research which appears in the form of *a priori*ty right. This novelty no doubt stems from the nature of the laboratories which participate in the operation; they are yeast biology laboratories and not merely "mercenary" sequencers. The possibility is thus created, before publication of the chromosome, of appropriating the sequence, either by scientific publication or by patenting. This system can also be seen as an experimental process in which different procedures can be tested. No attempt is made to decide in advance on the public or private nature of the sequences.

Conditions required by this type of approach

The effectiveness of the organization of yeast genome sequencing depends on the combination of three key elements:

- The yeast genome was small enough (12 megabases) for work to be allocated to laboratories with small "craft" sequencing capacities. The choice of the yeast genome cannot be dissociated from technical conditions in 1988-89. Sequencing techniques were in their infancy at that stage and sequencing itself constituted a bottleneck in many laboratories. Specialised databases and the exploitation of such information were barely advanced. Only homological comparison techniques (oligonucleotide and protein) were widely available. Sequencing was carried out by laboratories specialised in biology and not by large research centres since these did not yet exist at the time.
- Preliminary reports (Danchin, A., 1987) and articles seldom refer to debate on the choice of a complete sequencing for yeast[**]. Similarly, it seems that the decision to separate sequencing and functionalization was not debated; everything happens as if this option was tacitly ratified.
- Support from public authorities - European and national - was above all financial. The European communities funded the sequencing of the following chromosomes either fully

[**] Except perhaps the article by J. Ninio (Ninio, J., 1992) which compares the work of exploiting sequences to the exploitation of data in a telephone directory. However, Ninio quotes no author other than himself to support his argument.

or partially: II, III, IV, VII, X, XI, XII, XIV, XV, XVI. European laboratories also received national funds (Cf. Acknowledgement (Dujon B.e.a., 1994; Oliver S.e.a., 1992)).

This additional funding is not taken into account when the cost of sequencing the yeast genome is estimated. Although the variety of modalities of additional funding make it difficult to obtain reliable figures, it appears that such funds are equivalent to between 50% and 100% of EEC funds.

Table 1 : Cost of sequencing

Ch.	Coordinat.	size in Kb	co-ordinator in Kecu	Mips in Kecu	sub-contracting in Kecu	overlap and verificat.	Add. funds by national authorities	Total EU and nat. funds
2	Feldmann	807	80,7	30	1 614		920,6	
3	Oliver	314	47,6	156	1 570		874,3	
4 L2	Jacq	600	30,0	15	1 200		684,4	
7	Tettelin	1 150	114,0	15	2 300		1 311,8	
10	Galibert	720	72,0	15	1 440		821,3	
11	Dujon	666	66,6		1 332		759,7	
12 L	Hoheisel	450	22,5	15	900		513,3	
14	Philippsen	810	81,0	15	1 620		924,0	
15	Dujon	1 150	114,0	15	2 300		1 311,8	
16 L2	Goffeau	300	15,0	15	600		342,2	
	fixed costs			21				
	Total E.U.	6967	643,4	312	14 876	2 050,9	8 463,5	26 345,8
	USA	2 500						
	UK	2 190						
	Canada	535						
	Japan	270						
	TOTAL	12 462						

estimation of costs:
-sub-contractor : 2 Ecus/base except BAP with 5 Ecus
-verification and overlaps : 658 Kb overlaps intra EEC and 671 Kb diverse verifications at 1 ECU
- estimation of additional funds by national authorities: 50% of costs financed by EEC on the basis of 2 Ecus per pair of bases (rough estimation considering the different types of funding - e.g. purchase of material, but not researchers salaries).

The total sequencing costs are around 26 245 Kecus for 6967 Kb, i.e. 3,781 Ecus per base (4,79 USD). To the sequencing costs must be added functionalization costs (EEC Eurofan contract for 7320 Kecus). In total, the sequencing and functionalization budget of yeast genes is around 35 million Ecus for the 6967 Kb, which places the cost of the 3800 genes sequenced by the E.U. at between 10,0 and 12,6 Kecus (and between 24,1 and 30,6 Kecus for 1400 new genes). These figures are higher than estimates made by A. Goffeau and A. Vassarotti (Goffeau, A. and Vassarotti, A., 1993) (20 million ECU) who take into account neither national funds nor functionalization costs.

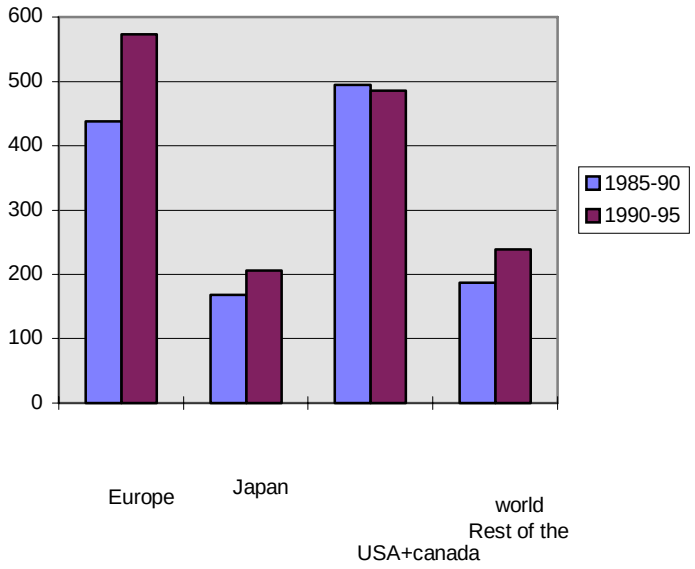
Impact of European sequencing programmes

Apart from the complete sequencing of the yeast genome - an undeniable scientific achievement - the distribution of work among the numerous European laboratories has been instrumental in structuring of the yeast research community in Europe. The exploitation of information from bibliometric databases provides some indication of these effects (Cf. Box 1).

Closer attention focused on Europe

Diagram 2 demonstrates the upsurge of Europe after 1990, when the E.C. firmly supported the sequencing of the yeast genome.

Diagram 2 : Respective evolution of "yeast genome" scientific production in various parts of the world between the periods 1985-90 and 1990-95.



Source : DBA database

The results are highlighted by the fact that Europe is the only area in which scientific production has soared, while growth in Japan is slow and uniform and the US and Canada show a slight decline.

Yeast, a model organism

The use of studies on yeast as a model for sequencing the human genome is often put forward by researchers to justify their interest in this micro-organism. In order to identify how and by whom yeast is used as a model, we used the BCI database to compile six files on the 1000 most productive and most cited authors on the genome (in general) and then on the yeast and plant genomes. Our analyses revealed:

- limited relations between the different genomes (4,4% of the authors were in all three files);
- 148 authors common to the yeast and general files (160 for the plant file). Out of these 148 authors, over 50% are productive on the human genome.

Thus, it seems that yeast is indeed used as a model, particularly by researchers who work on the human genome, for testing new methods or tools on a micro-organism before "trying" it on humans.

2. CONDITIONS FOR EXPORTING THIS MODEL OF COOPERATION

Specific conditions related to the organism

The detailed organization of yeast sequencing under the aegis of the EEC cannot be dissociated from a given scientific and technological context: reduced size of the genome; fledgling sequencing techniques; sequencing carried out by yeast research laboratories; important budgets and widespread scientific agreement on the programme to be conducted. When a programme is successful there is a strong temptation to apply the same methods to similar problems. In particular, we recall the sequencing of Arabidopsis thaliana. Yet, between Saccharomyces cerevisiae and Arabidopsis thaliana, conditions are radically different; the Arabidopsis thaliana genome is ten times bigger that that of yeast and sequencing has become a routine, industrial activity. Hence, "craft" sequencing is no longer an economically viable organizational modality, although resorting to a limited number of laboratories which already have significant sequencing capacities is politically tricky since it favours certain countries which already enjoy a privileged position.

During the sequencing of the yeast genome, potential actors were easily identified and most of them were already involved in sequencing. For Arabidopsis thaliana, on the other hand, the identification of actors was more difficult since technical conditions introduce a

necessary separation between sequencing and the use of sequencing. Indeed, very few plant laboratories were able to sequence effectively. It seems, moreover, that the Arabidopsis thaliana genome presented more intrinsic difficulties than did the yeast genome.

Exact duplication is therefore problematical. Yet, the principles of payment by the piece and standardisation of results can easily be retained. By contrast, given the distinction between sequencing laboratories and user laboratories, and the absence of unanimous scientific agreement on tasks to be undertaken, Principle 3 - the most innovative - will have to undergo substantial amendments before it can be applied elsewhere.

Evolutions of organization are required

Complete sequencing of the Saccharomyces cerevisiae genome generated numerous controversies between French and American teams; the latter criticised the former for not immediately making the results of sequencing available to the scientific community at large. The Europeans replied that the production of clean and ordered sequences had a high added value that they alone provided, and that this type of procedure needed time. Even within European teams, the MTA (Material Transfer Agreements) accompanying the exchange of genetic material generated widespread controversy when these documents specified conditions for use that were too restrictive, or unacceptable citation requirements (the need systematically to feature in second position on the list of authors of all articles in which the material was used). These two points illustrate appropriation problems which must be taken into account for sustainable co-operation.

In emergent scientific fields such as genome research we find the circulation not only of articles and patents, but also of competencies embodied in persons, in experimental devices and in genetic material. The status of everything circulating between the different laboratories and researchers, whether embodied in humans or in artefacts, is uncertain. It may concern research-related material, unpublished preliminary results, or simply genetic material (e.g. purified proteins).

Studies on genetics and molecular biology carried out by sociologists of science have clearly shown the role of access to data in scientific practice (Hilgartner, S., 1994; Hilgartner, S., 1995; Hilgartner, S. and Brandt-Rauf, S., 1994). Researchers develop real strategies to appropriate data, know-how, research material and intermediate results which are obtainable only at a huge cost (in terms of both time and money) and are difficult to

valorise. Drawing upon an analysis of the discovery of sex hormones at the beginning of the century, N.Oudshoorn (Oudshoorn, N., 1990) shows how the access to research material is a critical resource. Whoever has access to sufficient quantities of research material is a generator of scientific progress. Thus, access to data may constitute a source of « contributions in kind » when co-operation is being agreed:

> "I've got these clones, you've got expertise with this technique, let's pool our resources and do the following project", (Hilgartner, S. and BrandtRauf, S.I., 1994).

Yet, the uncertainties regarding their nature makes it difficult to take them into account in economic and legal terms. Depending on the moment or the place in which they are studied, they are either the results of research or inputs into research. Furthermore, the type or degree of originality of research material is completely dependent on the state of the science and of the networks to which the laboratories belong. For some of them access to mutants is problematical since they do not produce any themselves and they maintain only loosely coupled[4] relationships with those laboratories which do. For others, the access to mutants is not a problem since they are well integrated in the production networks. Access to contig cards could be, however, problematical. Thus, the more or less critical nature of resources to which the laboratory would like to have access depends on its position in the networks of other laboratories with which it works.

The strategies formed around data are increasingly diverse when these data are heterogeneous (intermediate results, research material, software, data bases, etc.).

Hilgartner et Brandt-Rauf (1994) identify three generic strategies for reducing access to data:

1. *Non release of data*: This strategy consists of keeping data private before the publication of the research. A typical case is that of a group which constructs a contig card from a public library. In this case the release of the contig card would be directly exploitable and could provide rival laboratories in the « gene hunt » with an advantage.

2. *Delayed access*: Proposing delayed access to data enables the producers of this data to maintain a big enough gap compared to rival groups. Once the laboratory has exploited the intermediary results there is no reason not to make them available to the entire scientific community; on the contrary, for if another group uses the data it will have to cite the laboratory and researchers that first produced it.

[4] (Weick, K., 1991).

3. *Data isolation*: This concerns the provision of access to partial data which cannot be used directly by another laboratory. By withholding a part of the information the rest cannot be used to reproduce the experiment or innovation. The advantage for the laboratory lies in making its presence known and in showing its lead so as to discourage potential competitors.

Yeast sequencing in Europe was mainly based on delayed publication even if the status of data was not clearly defined. *De facto* this contractual system created a continuum between the public and private status of data[5] (team data, pooled data, quasi-public data, public data).The play on these different levels of status allowed for compromise between incentives to research, the co-ordination of teams, and the development of partnership. For a year a sequence was reserved for a laboratory for its biological research (team data); it could be tested « blindly » by other laboratories in the consortium (pooled data); and, lastly, information on the sequence was released to the Yeast Industrial Platform (quasi-public data).

CONCLUSION

Such an organization underlines several problems and shows what kinds of modifications are required :

- the main problem is the cumulativeness of science. When publication of sequences is delayed, circulation of knowledge and research materials is also delayed as well as scientific progress. Whereas the problem is simple, the solution is not. It is necessary to find a system which allows private appropriation of scientific production and which encourages knowledge and data (research materials, partial and incomplete results, etc.) circulation. Delayed publication does not seem to be the right system for that.
- Faced with the scientific and institutional originality of the problems with which they are confronted, the actors invent new modes of appropriation founded on amended publication rules or on the creation of new contracts (the MTA). These new modes of appropriation emerge spontaneously to solve isolated problems of co-ordination, and are progressively generalised when adopted by other groups and researchers.

[5] We find here the same categories as those identified by Cassier in his analysis of the joint research programme on lipases (see Cassier, 1996, and Cassier and Foray, 1996).

- To be sustainable, co-operation in emerging scientific fields such as genomics needs a creation of new modes of organization of research which proposes a better balance between diffusion of knowledge and appropriation.

Box : The Databases

Longitudinal analyses have been carried out, drawing upon articles available in the Derwent Biotechnology Abstract (DBA) database available on CD-Rom since 1982. This is a practical base since it has been available on CD-Rom from the outset. On the other hand, its coverage is not as extensive as that of the Biotechnology Citation Index (BCI) which has been available on CD-Rom only since 1991. Analyses of citations and partnerships have been made on data in this base. 50903 references were extracted from the BCI, enabling the identification of 106190 different authors. The plant file contains 6033 references while the yeast file has 3342.

BIBLIOGRAPHY

Danchin, A., (1987), "Complete genome sequencig : futures and prospects" , Brussels: Commission of European Community, 1987

Dujon, B.; *et al.*, (1994), "Complete DNA sequence of yeast chromosome XI," Nature, 369, 371-378.

Feldmann, H.; *et al.*, (1994), "Complete DNA sequence Yeast of chromosome II," EMBO Journal, 13/24, 5795-5809.

Goffeau, A. & Vassarotti, A., (1993), "L'europe analyse le génome de saccharomyces cerevisiae," Biofutur, Novembre, 33-39.

Hilgartner, S., (1995), "Data access policy in genome research," Cornell University working paper, , .

Hilgartner, S., (1994), "The Human Genome Program", Handbook of science and technologie studies,. Ed. T., J.S.M.G.P.J.P.: SAGE.

Hilgartner, S. & Brandt-Rauf, S., (1994), "Data access, ownership and control," Knowledge: Creration, diffusion, utlisation, 15/4, 355-372.

Hilgartner, S. & BrandtRauf, S.I., (1994), Controlling data and resources : acces strategies in molecular genetics, CEPR/AAAS conference, Stanford,USA.: , 35.

Ninio, J., (1992), "Une biologie de retardataires," médecine/science, 8, 374.

Oliver, S.; *et al.*,, (1992), "The complete DNA sequence of yeast chromosome III," Nature, 357/7 May 1992, 38-46.

Oudshoorn, N., (1990), "On the making of sex hormones: Research materials and the production of knowledge," Social Studies of Science, 20, 5-33.

Vassarotti, A., *et al.*, (1995), "Structure and organization of the European Yeat Genome Sequecing Network," Journal of Biotechnology, 41, 131-137.

Weick, K., (1991), "The nontraditional quality of organisation learning," Organization Science, 2/1, 116-124.