From genome to interactome and beyond

Johan Hoebeke [1] .

[1]Centre National de la Recherche Scientifique hoebeke@ibmc.u-strasbg.fr

Correspondencia: Instituto de Medicina Tropical - Facultad de Medicina - Universidad Central de Venezuela.

Consignado el 10 de Julio del 2006 a la Revista Vitae Academia Biomédica Digital.

RESUMEN

La visión reduccionista que presentaba a la genómica como la única herramienta capaz de explicar las manifestaciones fenotípicas de los organismos así como su evolución, resultó estar equivocada. Si bien la genómica constituye una base fundamental para entender los mecanismos subyacentes a la vida, no es por sí sola capaz de explicarlos. La transcriptómica y la proteómica han revelado la complejidad de la expresión del código genético, pero aún no logran determinar el papel de las proteínas en la expresión del fenotipo. En ese sentido, es necesaria una visión más integrativa que permita analizar la función de las proteínas tomando en cuenta el metabolismo celular y el ambiente intracelular en el que se desenvuelven. La aparición de nuevas técnicas de biofísica y biología molecular ha permitido examinar las interacciones entre dos o más proteínas (interactómica), agregando una nueva dimensión al estudio complejo de la biología celular.

KEY WORDS: Genomic, interactome, proteomic, transcriptome

FROM GENOME TO INTERACTOME AND BEYOND

It is now thirty years ago when Richard Dawkins published his controversial "The selfish gene" (1) in the heydays of DNA sequencing and the start of the Human Genome Project (HUGO). According to the promoters of the project and in line with the thesis of Richard Dawkins that replication and conservation of genes were the driving force of evolution, knowledge of the DNA sequence of all human genes would unravel the secrets of the book of life. Biology would reduce itself to molecular biology *sensu strictu* and disciplines like systematics, physiology, anatomy, etc.,

would disappear as superannuated. Gene therapy, based on our knowledge of gene deficiencies in genetic diseases as well as in cancer or cardiovascular diseases, would be the new panacea. While DNA sequencing has vastly enlarged our biological knowledge, it has also highlighted the fallacies on which Dawkins and molecular biologists based their claims.

The Dawkins' fallacy is easy to handle: proposing the conservation and replication of genes as the driving force for evolution corresponds to the hypothesis that stones are the driving force in the evolution of architecture. Without expression of phenotypes, no Darwinian evolution is possible similarly as adaptation of housing to different climatic conditions lies at the basis of architectural changes and not stones or bricks, although these are essential to the construction of the house as the genes are essential to allow phenotypic expression. To use another metaphor, pretending that gene sequencing might unravel the book of life corresponds to the claim that knowledge of the Cyrillic alphabet is enough to understand Dostoyevsky's novels. While this is an essential requisite, it is far from sufficient to grasp the authorís ideas.

The wonders expected of gene therapy were based on another form of reductionism i.e. the simplistic idea that insertion of a deficient gene might be sufficient to redress the diseased phenotype. What was forgotten is the fact that expression of the inserted gene will depend on the place where it is inserted in the genome and the environment in which the insertion is made. The first omission led to tumour development in children successfully treated for a genetic immune deficiency (2); the second omission led to the death of a patient treated with a too high dose of the virus in which the deficient gene was inserted (3). Finally, it became clear that the Watson and Crick dogma of one gene, one protein was to be abandoned in view of the abundant splicing variants demonstrated in a majority of genes, making the rigid concept of gene fuzzier. This probably explains the surprisingly small amount of genes found in the primates compared to the wide variety of proteins (e.g. a few hundred genes account for the billions of potential antibodies).

A humble reassessment of what DNA sequencing has learned us is shown in Figure 1A, in which the letter sequence hides a philosophical sentence. Careful analysis can help us to extract at least the words used in this sentence; it will never be possible to get to the sense of the sentence. In a similar way, bioinformatics can help us to extract the words without giving us the meaning of the sentence. Other tools are therefore required.

A: genome	ACCGGTACTAAGGTCGGTAAATTTCGGGG GAAATTTT	tabheacvghbd nabcaojggforbnm oaigincuvbratreccvbduate uutopqhwothreetendghnt hobghusacnccghdabikleibbicngs

Figura 1. Metaphoric illustration of the molecular biological approach

A first breakthrough was the implementation of macroarrays, using cDNA clones obtained as bacterial colonies or PCR products, arrayed on Nylon membranes and hybridised with radioactive samples prepared by reverse transcription of total or messenger RNA (4). Miniaturisation of this technique led to the production of DNA chips, relying on the use of a large number of relatively short oligonucleotides synthesised in situ to detect transcripts, fluorescently

labelled by reverse transcription in the presence of modified nucleotides (5). These techniques allowed leaping from a static knowledge of the genome to the expression of the gene as an mRNA, the intermediate between gene and protein.

The use of DNA chips led to a rapidly increasing literature, pretending to link the expression of one or several genes to disease states or to cell dysfunction with the promise to find new therapeutically useful targets. The first enthusiasm was however dampened by the technical and the biological limitations of the method. The technical limitations are the unknown quantities as the exact size and shape of the probe spots, the density of probe DNA in each spot, the hybridisation efficiency and the labelling efficiency of a given sequence.

While these technical limitations can be overcome by the multiplication of the number of experiments, the multiplication of probes for the same gene and the use of differently labelled nucleotides on the same probe, thus increasing the time and the cost of a differential transcriptome determination, the biological limitations are more difficult to tackle. Indeed, a microarray is just a comparison of a particular gene expression in two RNAs with the inherent variability in space and in time of all biological material. This inherent variability can only be assessed by increasing the number of replications of RNA sample selection, which is often quite difficult using small amounts of biopsies composed of different types of cells or, in the case of cells in culture, to obtain samples at exactly the same period of the cell cycle.

Finally, there is a logical gap that has to be resolved. Is there a causal relation between the expression or the repression of gene transcription and the functional importance of the expressed or repressed gene? The answer is: as yet we do not know and if so, the causal relation has to be very tenuous in view of the general discrepancy observed when gene expression is compared at the transcriptome level with expression at the proteome level (6).

Another method to study the phenotypic importance of a gene is to inhibit transcription of the gene by knock-out (7) or iRNA silencing (8). A large number of the conclusions drawn by these methods however suffer from a logical flaw. As Aristotle proposed in his logic, a deduction with an affirmative conclusion must have two affirmative premises. From the absence of a gene, it can thus not be inferred that the gene is directly responsible for the wild type phenotype. Similarly, from the absence of a knock out phenotype, it cannot be inferred that the gene plays no role in the wild type genotype. Both fallacies take into account neither the complex regulatory genetic and epigenetic mechanisms leading from a genotype to a phenotype, nor the redundancies allowing for compensatory mechanisms upon a genetic deficiency.

To summarise as illustrated in Figure 1B, the transcriptome has facilitated the lecture of the words used in our philosophical sentence but the general sense remains hidden.

B: ACCGGUACUAAGGUCGGUAAAU dao thousand give be origin one two create transcriptome UUCGGGGGAAAUUUU three ten

Figura 1. Metaphoric illustration of the molecular biological approach

While protein chemistry lies at the origin of biochemistry, the discipline was partially eclipsed by molecular genetics after the discovery of the Watson - Crick Model, the unravelling of protein

biosynthesis and the use of restriction enzymes to isolate genes for transcription and translation (biotechnology). One of its handicaps was the necessity to obtain large amounts of material needed to analyse protein primary structure by Edman sequencing and protein tertiary structure by X-ray diffraction analysis, restricting advances in the field to proteins abundant in nature. Miniaturisation of protein sequencing, the identification of minute amounts of proteins by combination of 2D-electrophoresis and mass spectrometry and the possibility to use biotechnology to produce larger amounts of proteins has released protein chemistry from its initial handicap. A new era of proteomics started (Figure 2).

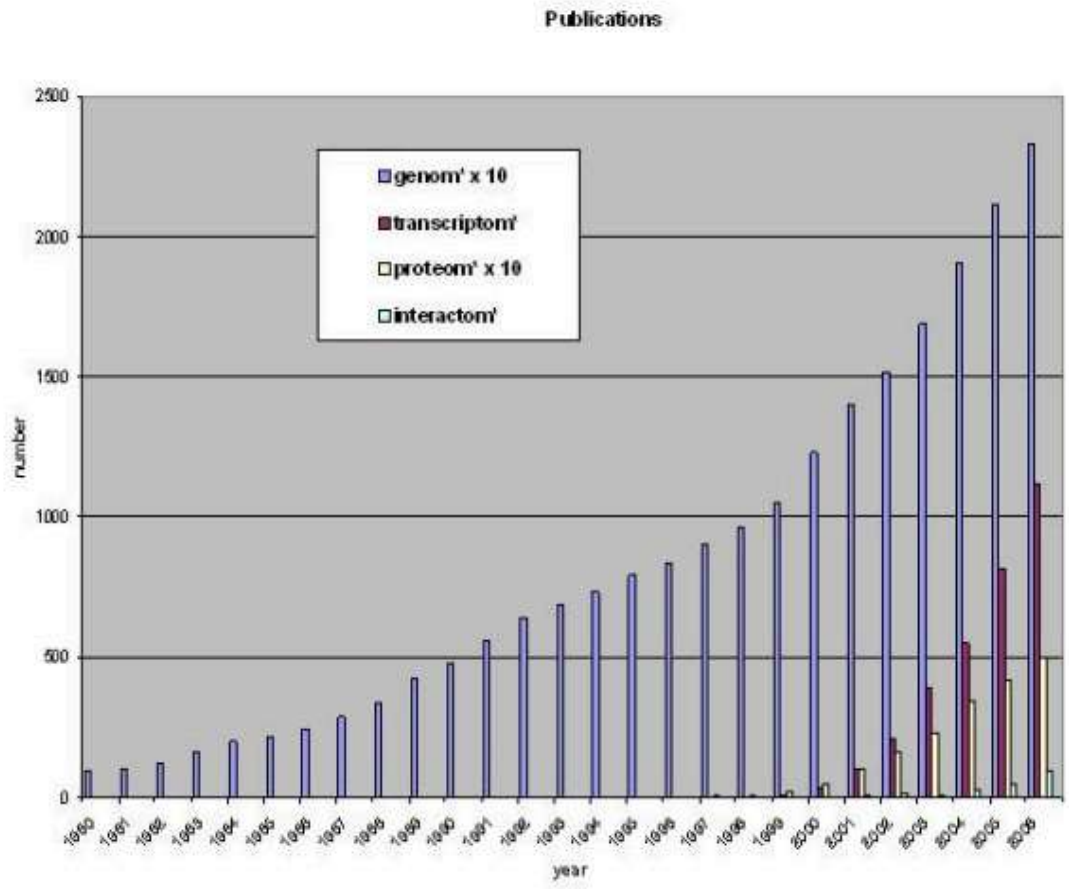


Figura 2: Number of publications in PubMed

The use of mass spectrometry coupled to 2D-electrophoresis or to capillary chromatography allows the study not only of protein expression but also of the post-translational modifications, which are a general feature of protein expression. While the technical limitations of proteomic analysis are few, the biological limitations of inherent sample variability in space and time remain. As was recently stressed, the dynamics of protein turnover are a missing dimension in proteomics (9). The logical gap mentioned earlier also remains since there is no direct correlation between phenotype and increase of decrease of protein amount. Indeed, the phenotype is the result of a protein network in which the protein at the lowest concentration or the enzyme with the lowest catalytic activity will be predominant in a cascade leading to the phenotype. An increase or decrease of at least twice or thrice (sensibility of the proteomic approach) might thus be less relevant than the increase or decrease of a few percent (under detection limit) of the rate limiting protein. Even if the sensibility of the method increases, the noise of biological variability

will preclude the detection of this protein. Other approaches are thus necessary in order to link protein synthesis to phenotype.

Identification of a protein, although important, needs a further description of its function in order to gain insight into its biological relevance. For a long time it was assumed that knowledge of the 3D-structure of a protein was sufficient to understand its function, initiating the paradigm of structure-activity or structure-function relationships. High resolution analysis of 3D structure by X ray diffraction seemed thus the Rosette stone to gain insight into biological processes at the molecular level. The structure-activity paradigm unfortunately has an inbuilt logical flaw. Activity cannot be defined without the notions of motion and time; structure is per definition spatially limited and is time independent. An increasing number of examples highlight this flaw : e.g., the interaction site between rhodopsin and transducin has no clear structure in the rhodopsin crystal, the structure of the epitope recognised by neutralising antibodies in the foot and mouth disease virus could not be resolved in the virus crystal.

We have to deal with the chemical extension of the physical uncertainty principle of Heisenberg: the simultaneous determination of velocity, or any related property e.g. energy or momentum, and position is impossible. To combine structure and activity, we thus need complementary spectroscopic techniques as NMR (nuclear magnetic resonance) spectroscopy, fluorescence spectroscopy, EPR (electron paramagnetic resonance) or ESR (electron spin resonance) spectroscopy and the use of molecular dynamics to model the obtained data in the structural framework derived from X-ray crystal diffraction. Recently, time-resolved X-ray crystallography has allowed to approach protein function at the nanosecond level (10), allowing for a comparison between molecular dynamics calculations and experimental results. Most of the protein association or enzymatic reactions however occur at the µs to s level, very difficult to reach by molecular dynamics in view of the calculation time needed. Cryo-EM permits to study motions at the ms level with the drawback that changes in conformation and viewing angles can be confounded and the extremely low signal-to-noise ratio (11). The Heisenberg principle as yet remains valid.

The field of proteome has now been enlarged to that of glycome, lipidome and metabolome respectively the study of oligo- and polysaccharides, the study of lipid environment and the study of metabolic pathways. To continue to use or syntactic metaphor, all these fields allow us to know how the different words corresponding to a protein are conjugated thus enabling us to get at least a rough idea about the meaning of the sentence (Figure 1C and D). To really get insight into the meaning of the sentency, we need to correctly order the different words (Figure 1E). The biological discipline trying to make order in the topology of the cell proteins is called: the study of the interactome.

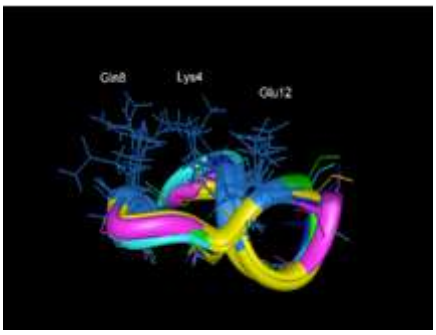
C: proteome	MADEKEGALLRK INVTGGP?..		dao thousand beings original gives one two creates three ten

Figura 1. Metaphoric illustration of the molecular biological approach

D : lipidome, glycome, metabolome	Associated glucids, lipids, cofactors etc.	the dao thousand beings original one two gives creates three ten to

Figura 1. Metaphoric illustration of the molecular biological approach

Until recently, due to technical limitations, interactions between two proteins were always biased in the sense that the investigator ought to choose the partners in advance before testing the hypothesis that they interact. The goal of interactome studies is to bypass the bias and to study the interaction between one or several partners in an undefined mixture of proteins. This new approach has been made possible by the introduction of two techniques: one derived from molecular biology and called the two hybrid technique (12), the other using surface plasmon resonance combined with mass spectrometry (13).

The first one is more powerful since it allows working with two or three indefinite partners; the second one has one defined partner which is allowed to interact with second partners present in a mixture of indefinite proteins. It is not the purpose here to describe the two techniques but to emphasize their advantages and drawbacks. The two hybrid techniques permit a high throughput screening of interactions between millions of potential interactions (14). The success-rate is quite high (~80%) but false positives are not negligible. A major drawback however is the uncertainty of the kinetics of the interacting proteins.

Both hybrid systems thus allow the construction of a 3D topology of proteins interacting in the cell but lack the fourth dimension which is time. In contrast, surface plasmon resonance allows kinetic analysis of the protein-protein interaction. It has however not the high throughput capacity of the two hybrid technology and does not allow the identification of fast dissociating proteins since coupling to the mass spectrometer necessitates a few minutes which might be a long time for short but functionally significant interactions. Moreover, the sensitivity is limited to relative low molecular weight interactants (< 30 kDa). The two methods are however complementary and can be verified by other techniques such as pull down immuno-precipitation and fluorescent techniques to analyse the interaction of the identified pair. As shown in Figure 2, the interactome study is at its beginning but is increasing exponentially.

Will the knowledge of the complete cellular interactome superannuate cell biology? The answer is illustrated in Figure 1E. Although the interactome allows us to propose an ?educated guess?

about cell biology, it will never replace the information obtained by the tools of cell biology. As it is necessary to learn the Chinese language to be sure that our guesses about the sentence are correct, cell morphology and physiology use different techniques with a different language to study the global cell structure and function which always will be larger than the sum of all molecular interactions. I started with the history of the genomic reductionism at the beginning of my article. Let us not stumble in the same reductionism at the end of it, forwarding the interactome study as the panacea to solve all biological problems.

E : interactome Interaction between proteins and environment

the original dao creates the one,
the one creates the two,
the two creates the three,
the three gives origin to
the ten thousand beings
Lao Zi

Figura 1. Metaphoric illustration of the molecular biological approach

REFERENCES

1. Dawkins R. (1976) The selfish gene. Oxford University Press, G.B.
2. Haceyn-Bey-Abina S., von Kalle C., Schmidt M., Le Deist F., Wulffraat N., McIntyre E., Radford I., Villeval J.L., Fraser C.C., Cavazzana-Calvo M. and Fischer A. (2003) A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N. Engl. J. Med.* **348,** 255-6.
3. Marshall E. (1999) Clinical trials: gene therapy death prompts review of adenovirus vector. *Science* **286**, 2244-45.
4. Lennon G.G. and Lehrach H. (1991) Hybridization analysis of arrayed cDNA libraries. *Trend Genet.* **7,** 314-17.
5. Schena M., Shalon D., Davis R.W. and Brown P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467-470.
6. Kuo C.C., Kuo C.W.,Liang C.M. and Liang S.M. (2005) A transcriptomic and proteomic analysis of the effect of CpG-ODN on human THP-1 monocytic leukaemia cells. *Proteomics* **5,** 894-906.
7. Zijlstra M., Li E., Sajjadi F., Subramani C. and Jaenisch R. (1989) Germ-line transmission of a disrupted beta-2-microglobulin gene produced by homologous recombination in embryonic stem cells. *Nature* **342,** 435-438.
8. Fire A., Xu S.Q., Montgomery M.K., Kostas S.A., Driver S.E. and Mello C.G. (1998) Potent and specific genomic interference by double-stranded RNA in *Caenorhabditis elegans. Nature* **391**, 806-811.
9. Pratt J.M., Petty J., Riba- garcia I., Robertson D.H., Gaskell S.J., Oliver S.G. and Beynon R.J. (2002) Dynamics of protein turnover, a missing dimension in proteomics. *Mol. Cell. Proteomics* **8,** 579-91.
10. Schotte F., Lim M., Jackson T.A., Smirnov A.V., Soman J., Olson J.S., Phillips Jr. G.N., Wulff M. and Anfinrud P.A. (2003) Watching aprotein as it functions with 150-ps time-resolved X-ray crystallography. *Science* **300,** 1944-1947.

11. **Berriman J. and Unwin P.N.T.** (1994) Analysis of transient structures by cryo-electron microscopy combined with the rapid mixing of spray droplets. *Ultramicroscopy* **56,** 241-252.
12. **Fields S. and Song O.** (1989) A novel genetic system to detect protein-protein interaction. *Nature* **340**,245-6.
13. **Zhukov A., Schurenberg M., Jansson O., Areskoug D. and Buijs J.** (2004) Integration of surface plasmon resonance with mass spectrometry: automated ligand fishing and sample preparation for MALDI MS using a Biacore 3000 biosensor. *J. Biomol. Tech.* **15,** 112-9.
14. **Rual J.F., Venkatesan K., Hao T., Hirozane-Kishikawa T., Dricot A., Li N., Berriz G.F., Gibbons F.D., Dreze M., Ayivi-Guedehoussou N., Klitgord N., Simon C., Boxem M., Milstein S., Rosenberg J., Goldberg D.S., Zhang L.V., Wong S.L., Franklin G., Li S., Albala J.S., Lim J., Fraughton C., Llamosas E., Cevik S., Bex C., Lamesch P., Sikorski R.S., Vandenhaute J., Zoghbi H.Y., Smolyar A., Bosak S., Sequerra R., Doucette-Stamm L., Cusick M.E., Hill D.E., Roth F.P. and Vidal M.** (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437,** 1173-78.