Using genomics to understand the nervous system

Claes Wahlestedt

Genomic approaches have the potential to affect almost every aspect of neuroscience. This review outlines recent advances in genome sequencing and genome variation research as well as the related genomic and bioinformatic technologies. Genomic strategies to study targets and pathways in the nervous system are discussed within the context of drug discovery efforts.

Claes Wahlestedt
Center for Genomics and Bioinformatics
Karolinska Institutet
SE 171 77 Stockholm
Sweden
tel: +46 87 28 66 29
fax: +46 83 23 95 0
e-mail:
claes.wahlestedt@cgr.ki.se

▼ Today, the vast majority of human genes, particularly those pursued in drug discovery efforts, have been identified *in silico*. As a consequence, we now have large numbers of potential drug targets that need to be prioritized and validated. Many drug discovery related genomics efforts are increasingly directed towards correlating findings in model organisms with human diseases. Although genotyping technologies are generally not yet as powerful as desired, scientists are now positioning themselves to understand the genetic basis of common complex disorders affecting the nervous system through studies of genomic polymorphism.

For a number of years, we have defined genomics as the scientific discipline of sequencing, mapping and analyzing genomes. The ever-growing analysis of aspects of genomics have often been used more or less synonymously with the term 'functional genomics', which, in turn, already overlaps to a great degree with many important lines of neuroscience research, and will increasingly do so.

Currently, genome sequencing groups are working hard to finish the human genome map, with some two-thirds of it remaining in draft form[1,2]. Significant efforts are also being made to produce finished, or gap-free genome sequences from two other experimental animals of great importance to neuroscience within the next five years – the mouse and the rat. Genome sequences for other multicellular organisms, such as *Caenorhabditis elegans*[3] and *Drosophila melanogaster*[4], have existed in a useful form for some time and have been of importance to neuroscience research. From a neuroscience perspective, the rat genome is clearly required because the rat has been used more extensively than the mouse as a model organism and it is unlikely that all long-established and relevant rat behavioral tests can be readily adapted to mouse experimentation.

Proteomics represents an important scientific discipline that is currently far less advanced than genomics. Proteins are the direct targets of the majority of drugs and therefore warrant much attention. However, it must be realized that, compared with the genome and methods available to study events at the DNA or RNA level, the proteome is far more complex and experimental methodologies are less powerful. For example, only a fraction of the *C. elegans* proteome can currently be analyzed using state-of-the-art 2D gel technology[5]. Hence, certainly for the next few years, it is likely that (near-) global studies at genome and transcript levels will create hypotheses that will need to be followed up by non-global ('one protein at a time') studies at the level of the proteome, and not vice versa.

Bioinformatics

Genomics researchers use DNA sequence data and computational systems to store, access and analyze data, constituting the rapidly growing field of bioinformatics. Bioinformatics also increasingly bridges the gap between a nucleotide sequence and a 3D protein product, that is, the field increasingly referred to as structural genomics or proteomics.

The past few years have seen intense efforts to mine the genomic databases *in silico*. These efforts have resulted in the discovery of many novel genes without known function, often (but not exclusively) based on nucleotide patterns characteristic of gene structure. Research efforts to ascertain the physiological or pathophysiological activities of the products from these newly discovered genes require a multidisciplinary approach. A significant amount of important information still resides in

1359-6446/01/$ – see front matter PII: S1471-1931(01)00166-5

DDT Vol. 6, No. 15 (Suppl.), 2001

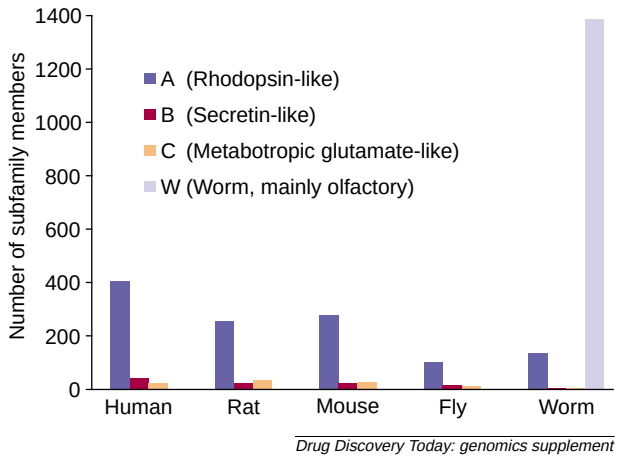

Figure 1. Known 7-transmembrane receptors (7-TM) present in the Swiss-prot and TREMBL databases in March 2001. Number of subfamily members are shown for the three main classes of 7-TM receptors and classes of olfactory receptors in worm.

the private domain, although many journals will only consider papers if authors make the sequences public. Databases with linear sequences of genomes, as well as databases of proteins and protein domains, are structurally well-developed, although many neuroscientists argue for the creation of shared databases containing multidisciplinary data from all aspects of neuroscience. Although the latter is important, it poses a major challenge to the field of neuroscience[6].

A gene family approach

Gene products within certain classes of genes have been proven to be 'drugable' and it is therefore often assumed that novel members of these families could be attractive drug targets in the future. In fact, the vast majority of current pharmaceuticals (excluding anti-infectives) target four broad classes of gene products: seven transmembrane (7-TM) receptors, ion channels, nuclear hormone receptors, and enzymes. Some 80% of targets for drugs used in psychiatry are 7-TM receptors.

The 7-TM gene family is very large, comprising up to 3% of all human genes, and the functions of many of the individual members of this family are completely unknown. Fig. 1 compares the number of known 7-TM receptors in human, rat, mouse, fly (*D. melanogaster*) and worm (*C. elegans*). Division of these receptors into sub-groups A, B, C and W illustrates the many similarities that exist between species with respect to 7-TM receptors[7,8], but also shows that a model organism approach can be misleading, for example if searching for mammalian homologs of worm olfactory receptors (Fig. 1).

Challenges in genomics

Great advances have been made over the past few years through *in silico* mining of databases. As already discussed, many candidate genes and targets in drug discovery have emerged, and prioritization efforts are now necessary. There is a need to gather much more information about the genetic basis of human complex diseases to follow up on relevant phenotypic information

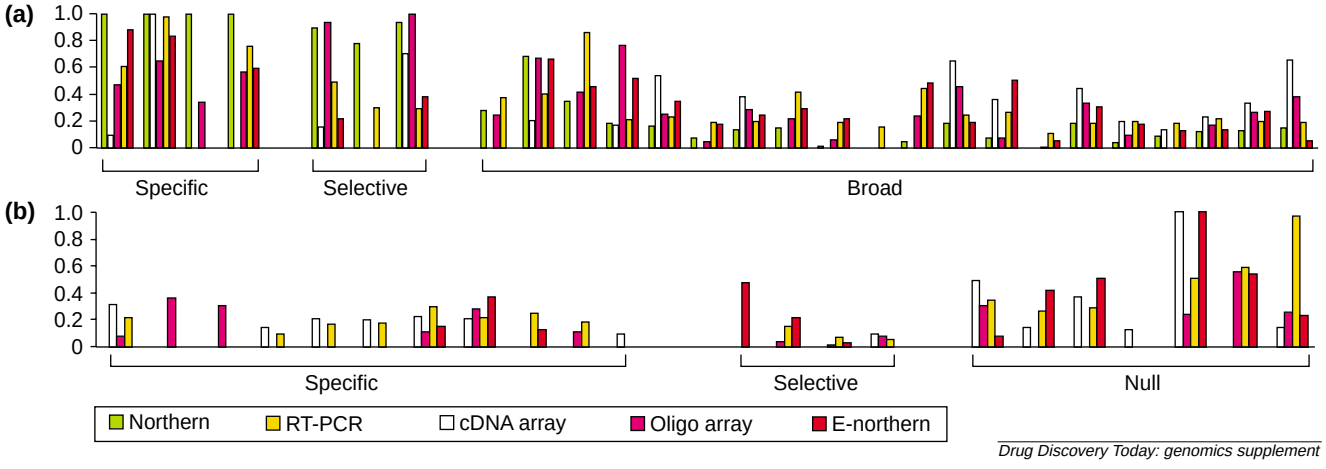

Figure 2. Varying results from different human tissue expression profiling techniques. Expression profiles were performed for 50 genes with five RNA samples from a commercial supplier (Clontech; Palo Alto, CA, USA). Genes were selected to represent broad human-protein families. Profiles were produced across the five tissues using reverse transcriptase-PCR (RT-PCR), oligonucleotide expression arrays (Affymetrix; Santa Clara, CA, USA) and cDNA expression arrays. In addition, data were collected from published northern blots that utilized RNA from the same supplier for the same five tissues. E-northerns were generated by analysis of the UniGene database (http://www.ncbi.nlm.nih.gov). The figure depicts the fraction of total expression observed in brain for (a) genes observed to have expression in brain identified by northern blots and (b) genes with no observed northern bands in the brain sample. 'Specific' indicates northern blots in which expression was only observed in one of the five tissues. 'Selective' indicates expression observed in only two tissues, 'Null' indicates no expression, and 'Broad' indicates expression across four or more of the tissues.

Table 1. Technologies and approaches for target prioritization and validation with *in vivo* focus

	Main advantages	Main disadvantages
Antisense oligonucleotide	Rapid, inexpensive, versatile and very effective in brain and spinal cord (and *in vitro*)	Uncertainty in the choice of chemistry and mRNA target site, incompleteness of effect
Ribozyme	No advantage over using antisense oligonucleotides	Low efficacy, biostability and cellular uptake
Mouse conventional knockout	Complete elimination of the gene	Slow, difficult, possible absence of detectable phenotype(s)
Mouse conditional knockout	Temporal and spatial control	Slow, difficult, incompleteness of effect
Viral vectors	Long duration of effect	Cumbersome, safety issues?
RNAi (double-stranded RNA)	High efficacy and easy to use	Can only be used for lower species
Engineered transcriptional regulation	Good efficacy (*in vitro*)	Lack of specificity? Unproven *in vivo*
Antibodies	Protein level targeting	Limited functional inhibition *in vivo*, labor intensive

in animals. Because of the lack of in-depth knowledge about human complex diseases, it follows that animal models of these diseases are not well developed. These challenges are not least apparent in neuroscience-related diseases that are of high importance to medicine and the pharmaceutical industry.

Prioritization and validation of drug targets in the nervous system

With the large number of newly discovered genes, it follows that any organization, small or large, needs to pursue prioritization of putative drug targets. Indeed, this is a bottleneck that must be adequately addressed in today's drug discovery efforts. Another important issue is that these targets can be polymorphic in humans (discussed later in this review), which increases the complexity of the problem further. The objective, therefore, often becomes to accumulate and analyze wet laboratory data in a high-throughput mode. This becomes particularly challenging when pursuing *in vivo* model systems, which involve the many complexities of the nervous system.

Microarrays

Much attention has been given to DNA microarray analyses over the past few years. Although these microarrays have enabled scientists to study large sets of genes in parallel, it has become increasingly apparent that this is only one tool among others that is applicable to functional genomics[9]. Many neuroscientists require systems with higher sensitivity and specificity, because minor, and often undetectable, changes in gene expression can be functionally important. As shown in studies of gene expression in the human brain (Fig. 2), oligonucleotide or cDNA microarrays do not provide as high sensitivity as quantitative PCR and/or radioactivity-based methods. Moreover, data analyses of microarray experiments often pose a major challenge because of the sheer magnitude of data generated and associated statistical issues. Further causes for concern include the problem of insufficient cross-validation procedures and quality control issues.

Other methods to monitor differential gene expression

There are a number of additional methods in use for transcript profiling. Among these, variations of differential display (DD), representational difference analysis (RDA), serial expression of gene expression (SAGE), and suppression subtraction hybridization (SSH), have all been applied to studies of the nervous system, resulting in the identification of many candidate genes. Generally, these findings must go on to be validated in other systems and further assessed for their relevance to human disease.

Regardless of which technique is being used to monitor gene expression, the rationale (i.e. the question being asked) and use of optimal control samples are both crucial issues. For example, gene expression analysis of the response *in vivo* to a particular drug might be meaningless when carried out shortly after administration (i.e. immediate, unspecific response), but could lead to the generation of new target hypotheses when carried out after long-term treatment in the animal model (i.e. specific changes in response to the drug given).

In vivo knockdown/knockout or overexpression

The most important pre-clinical aspect of the prioritization/validation bottleneck when studying the nervous system, is the production of experimental animals with altered expression of any particular gene product(s) of interest. Table 1 shows some

Table 2. Current understanding of the genetic basis of selected CNS disorders[a]

Disorder	Pattern of inheritance	Genes or loci
Early onset (familial) Alzheimer's disease	RAD	Three genes identified (presenilins 1 and 2, and amyloid precursor protein)
Late onset Alzheimer's disease	CC	Increased risk with apolipoprotein e4 allele; replicated
Parkinson's disease	RAD RAD or sporadic	α-synuclein parkin
Frontotemporal dementia	RAD or sporadic	tau
Huntington's disease	RAD dynamic mutation	Huntingtin (unstable trinucleotide repeat)
Familial amytrophic lateral schlerosis (ALS)	RAD	Superoxide dismutase (SOD-1)
Schizophrenia	CC	Many reported linkages, including chromosomes 1, 5, 6, 10,13, 15 and 22, but no replication
Attention deficit, hyperactivity disorder	CC	*DRD4* best replicated, others less certain
Fragile X mental retardation	Non-standard X-linked dynamic mutation	Two genes identified (*FMR1* and *FMR2*)
Dyslexia	CC	Two contributory loci on chromosomes 6 and 15; findings replicated

[a]Abbreviations: CC, common complex; CNS, central nervous system; RAD, rare autosomal dominant.

advantages and disadvantages associated with methods that can be used for 'knocking down' or 'knocking out' specific gene products, with respect to *in vivo* settings.

Antisense knockdown and related approaches

Among the higher-throughput and more drug-like approaches that create a situation *in vivo* where one or several genes are reduced in expression, the antisense approaches have been developed most extensively. The central nervous system (CNS) is perhaps particularly amenable to antisense oligonucleotide treatment because cerebrospinal fluid contains very low nuclease activities. Therefore, even unmodified DNA molecules can exert antisense effects[10,11]. Challenges in antisense research, and where significant advances have recently been made, include better targeting of mRNA and new chemistries with different properties[12].

Ribozymes are another tool that can be used instead of antisense oligonucleotides[13], but their use has several disadvantages, such as lower efficacy, lower cellular uptake, lower biostability and higher cost.

Transgenic mice

Manipulating the mouse genome is an approach that is extensively used in studies on drug targets. Over-expressing mouse transgenics can be used to study the phenotypic effects of gain-of-function or up-regulation of genes. In addition, gene knockouts in mice might demonstrate that loss of function of a particular gene is causative or contributory to the resulting phenotypic patterns. The latter might include observations concerning undesired consequences relating to the mechanism(s) that are involved. However, it is fair to say that many neuroscientists, in contrast to scientists studying less complex systems, have recognized the shortcomings of knockout mice as important and robust tools[14]. Specifically, an important issue is that conventional knockout mice show developmentally influenced phenotypes that can be misleading, and significant compensatory mechanisms probably exist in many cases. A way forward would be to produce more mice with conditional knockout strategies, where a gene can be deleted with temporal and spatial control. However, there are also disadvantages with these types of mice with respect to completeness of effect, and the time and high costs that are required for their production. Additional work is needed to improve conditional knockout mouse protocols.

Furthermore, viral vectors have also been used extensively for studies on the nervous system[15], although methods are still cumbersome. Viral vector approaches can provide long-term gene (or antisense) expression *in vivo* in, for example, the brain, and can therefore address issues related to sub-chronic or chronic alterations in protein level or pathway activity.

In non-mammals, there are other and often less complicated opportunities to pursue functional genomics. For example, the so-called RNA interference (RNAi) approach has been very successful in knockdown studies using *C. elegans*[16,17] and other non-mammalian species *in vivo*. RNAi relies upon the ability of double stranded RNA (dsRNA) to degrade the corresponding mRNA species. In worms, this convenient method often implies dsRNA delivery to the worms by their feeding on a particular

strain *of Escherishia coli* expressing the dsRNA of the gene of interest, including genes expressed in the nervous system[16,17].

Human genetics and disease

Our understanding of the genetic basis of neurological and behavioral disorders in humans is very much incomplete (Table 2 gives examples of disorders where intense efforts have resulted in important findings). In particular, common complex disorders such as schizophrenia are poorly understood with respect to genetics[18]. The replication of findings in different laboratories and cohorts is necessary to support any claims made concerning genes that are associated with increased risk of complex disease. Indeed, the vast majority of such claims that have been made have subsequently not been replicated. However, one well-confirmed finding is the increased risk that is associated with the apolipoprotein E4 allele in Alzheimer's disease[19].

Genetic epidemiology is currently, at least in part, shifting from meiotic (linkage) mapping in family-based studies to population-based association studies in unrelated individuals, exploiting high-density physical maps based on ordered single nucletoide polymorphisms (SNPs) and new statistical tools for analyses. In many current human association studies, candidate genes are prioritized based on results from human disease linkage studies and, ideally, also from biological knowledge.

SNPs, which by definition occur in at least 1% of the population and account for more than 95% of genetic variability in the human species, are the subject of intense focus, and large databases of SNPs are currently being generated (e.g. dbSNP and HGBASE). This work is initially based on DNA sequencing and the discovery of SNPs. Although some SNPs affect the coding part of the genome (cSNPs) and protein function in a direct way, the vast majority of SNPs are 'silent', yet remain useful as 'signposts'. Attempts to score (or test) these SNPs, ideally genome-wide using a large number of individuals (which is currently not possible), will push the boundaries of genotyping technologies. However, future improvements in genotyping technologies and increasing the number of SNPs tested can simply lead to an unmanageable overload of false-positive signals, thereby obscuring true disease-associations.

Perhaps the most valuable information about a novel gene follows once variant forms of the gene are linked to some human disease process. In drug discovery, diseases of interest typically involve complex etiologies, implying that 'risk modifying' rather than 'causative' sequence variants will be of interest. Positive findings will sometimes (but not always) support molecular hypotheses concerning the pathological basis of disease. However, when this is the case, the information can provide valuable evidence for the involvement of a putative drug target or pathway in human disease, not least those that affect the nervous system.

Acknowledgements

I would like to thank many of my colleagues at the Center for Genomics and Bioinformatics, and Pharmacia for valuable input. I would particularly like to thank Erik Sonnhammer and Wyeth Wasserman for help with Fig. 1 and Fig. 2, respectively.

References

1 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

2 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351

3 Consortium, T.C.e.S. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–2018

4 Adams, M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195

5 Schrimpf, S. *et al.* (2001) A two-dimensional protein map of *Caenorhabditis elegans*. *Electrophoresis* 22, 1124–1232

6 Chicurel, M. (2000) Databasing the brain. *Nature* 406, 822–825

7 Bateman, A. *et al.* (2000) The Pfam protein families database. *Nucleic Acids Res.* 28, 263–266

8 Remm, M. and Sonnhammer, E. (2000) Classification of transmembrane protein families in the *Caenorhabditis elegans* genome and identification of human orthologs. *Genome Res.* 10, 1679–1689

9 Luo, Z. and Geschwind, D.H. (2001) Microarray applications in neuroscience. *Neurobiol. Dis.* 8, 183–193

10 Wahlestedt, C. *et al.* (1993) Modulation of anxiety and neuropeptide Y-Y1 receptors by antisense oligodeoxynucleotides. *Science* 259, 528–531

11 Wahlestedt, C. *et al.* (1993) Antisense oligodeoxynucleotides to the NMDA-R1 receptor channel protect cortical neurones from excitotoxicity and reduce focal ischemic infarctions. *Nature* 363, 260–262

12 Wahlestedt, C. *et al.* (2000) Potent and nontoxic antisense oligonucleotides containing locked nucleic acids. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5633-5638

13 Salmi, P. *et al.* (2000) Dopamine D2 receptor ribozyme inhibits quinpirole-induced stereotype in rats. *Eur. J. Pharmacol.* 388, R1–R2

14 Crabbe, J.C. *et al.* (1999) Genetics of mouse behavior: interactions with laboratory environment. *Science* 284, 1670–1672

15 Bjorklund, A. *et al.* (2000) Towards a neuroprotective gene therapy for Parkinson's disease: use of adenovirus, AAV and lentivirus vectors for gene transfer of GDNF to the nigrostriatal system in the rat Parkinson model. *Brain Res.* 886, 82–98

16 Hsieh, J. and Fire, A. (2000) Recognition and silencing of repeated DNA. *Annu. Rev. Genet.* 34, 187–204

17 Vaz Gomes, A. and Wahlestedt, C. (2000) Altered behavior following RNA interference knockdown of a G-protein-coupled receptor in *C. elegans* using dsRNA delivery by feeding. *Eur. J. Pharmacol.* 397, R3–R5

18 Baron, M. (2001) Genetics of schizophrenia and the new millennium: progress and pitfalls. *Am. J. Hum. Genet.* 68, 299–312

19 Strittmatter, W.J. and Roses, A.D. (1996) Apolipoprotein E and Alzheimer's disease. *Annu. Rev. Neurosci.* 19, 53–77