Open-System Approaches to Gene Expression in the CNS

J. Gregor Sutcliffe

Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037

Where we started

In the present era of the nearing completion of the nucleotide sequences of the human genome and of several model organisms, it is easy to overlook the common viewpoint of 20 years ago, especially among neuroscientists. It was clear at that time intellectually that the genome encoded the protein set and that the proteins provided the hardware for the biochemical operation of the organism. Nevertheless, it was not widely evident that one would be able to determine the protein set via nucleic acid analyses in the relatively near term. In part, this belief was attributable to the vastness of the genome, the then-recently discovered fact that most protein-coding regions in the genome are interrupted by noncoding introns, and the lack of sufficient computing power to store and analyze the information. But it was also attributable to a generally held antipathy toward description-based studies: one cloned and sequenced genes whose protein products had been found already, either through biochemical or genetic studies, to be functionally interesting. Descriptive studies of the sort that are presently classified under the rubric of “genomics” were unfashionable (cf. Barnstable et al., 1983).

It was also not clear that even if one had the protein set as a list of putative amino acid sequences that this would give one much of a running start for understanding how any organ functioned, especially one as complex as the brain. The finding that a large percentage of proteins fall into families that share structures and biochemical activities has given instant meaning to many newly determined amino acid sequences. The advent of methods to produce synthetic and recombinant proteins to serve as biochemical and immunological reagents has greatly aided in the functional characterization of these newly discovered proteins, as have methods to manipulate their genes in experimental animals so as to alter their expression and activity *in vivo*. This article intends to show how conceptual and technological advances in molecular biology have moved neuroscience into the postgenomic era.

cDNAs represent the mRNA set

One advance of enormous importance was the development of techniques for producing cDNA and cDNA libraries. cDNA is mRNA that has been copied into DNA by the enzyme reverse transcriptase. We realized that cDNA libraries represent all of the mRNAs expressed in the tissue from which the sample was isolated and, thus, that such libraries could inform us about the complete protein set. By analyzing the size, abundance, and

tissue distribution of the mRNAs corresponding to nearly 200 clones isolated randomly from a rat brain cDNA library, Milner and Sutcliffe (1983) calculated that the 10^8 to 2×10^8 nucleotides of mRNA expressed by the brain corresponded to 20,000–40,000 distinct mRNAs. Of these, ~65% were enriched in the brain compared with peripheral tissues. Most were of low abundance, on the order of one part in 10^5 . The recently published draft sequences of the human genome (Lander et al., 2001; Venter et al., 2001) suggested that the total number of human genes is in the range of 30,000, but a more recent study based on matching cDNA sequences in databases to the genome sequence reassessed this to 65,000–75,000 (Zhuo et al., 2001).

This early study represents the beginning of what have since come to be known as open-system approaches to mRNA expression analysis: mRNAs are detected because of their property of being expressed in the tissue sample isolated for study. This approach is in contrast to what are called closed-system approaches, in which the mRNAs to be detected are selected as candidates in advance of analysis; a contemporary example of a closed system would be a gene “chip” hybridization experiment.

Detecting brain-specific mRNAs by subtractive hybridization

Which of the 20,000–40,000 mRNAs deserved in-depth characterization? At the time, sequencing technology had not been automated. Therefore, procedures for triaging cDNA clones were necessary. Initially, brute force screening for brain specificity was used. Because a substantial portion of the mass of mRNA in the brain corresponds to a relatively small number of highly abundant, ubiquitously expressed species, it soon became apparent that this would be a glacially slow approach for determining which among the 13,000–26,000 brain-enriched mRNAs were especially important in directing the many unique functional processes that the brain orchestrates.

The issue was one of throughput, and it has been addressed technologically on several levels. One approach has been to enrich cDNA libraries, via subtractive hybridization, for clones of mRNAs that are expressed with some degree of spatial or temporal specificity. This methodology, originally developed by Timberlake (1980) for studies on gene expression in fungi, has been progressively improved in the ensuing decades to a degree that it has allowed identification of mRNAs selectively expressed within complex mammalian nervous tissue: examples from this laboratory include retinal photoreceptor-specific mRNAs, one of which was the product of the mouse *retinal degeneration slow* (*rds*) gene implicated in hereditary retinitis pigmentosa (Travis et al., 1989); forebrain-enriched mRNAs, including RC3/neurogranin, the calmodulin-regulating phosphoprotein of dendritic spines (Watson et al., 1990), and cortistatin, the sleep-promoting, acetylcholine-antagonizing neuropeptide of cortical interneurons (de Lecea et al., 1996); striatum-specific mRNAs (Usui et al.,

These studies were supported in part by National Institutes of Health Grant GM32355. I warmly acknowledge the many collaborators who have shared this neurogenomic journey, especially Rob Milner and Floyd Bloom, who were there at the beginning, and my collaborators at Digital Gene Technologies, who automated the TOGA process and engineered its informatics.

Correspondence should be addressed to J. Gregor Sutcliffe, Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037. E-mail: gregor@scripps.edu.

Copyright © 2001 Society for Neuroscience 0270-6474/01/218306-04\$15.00/0

1994), including several components of the intracellular G-protein transduction system; and hypothalamus-specific mRNAs (Gautvik et al., 1996), including that which encodes the precursor of the hypocretin peptides (de Lecea et al., 1998), which are part of a complex circuit that integrates aspects of energy metabolism, cardiovascular function, hormone homeostasis, and sleep/wake behaviors. The human sleep disorders collectively termed narcolepsy result from insufficiencies in the hypocretin signaling system.

Despite the obvious power of these refined subtraction methodologies for identifying mRNAs that have a particular selective pattern of expression, the overall patterns emerged one gene at a time. The process is single minded, allowing only one distribution dichotomy to be queried per experiment. Thus, tedious follow-on studies are necessary to elucidate the expression pattern of each new mRNA. What were required were procedures that would provide a survey of gene expression so that data from several anatomical and behavioral paradigms could be simultaneously collected.

Expressed sequence tags

With the development of automation for DNA sequencing and computer databases for archiving and analyzing the sequence data, new open-system strategies for cDNA analysis emerged. It was now possible to obtain fragments of sequence from hundreds of randomly selected cDNA clones rapidly (Adams et al., 1991). Because cDNA represents the mRNA set, these short sequences were dubbed expressed sequence tags (ESTs). The initial studies were only a small step beyond those of a decade earlier. However, sequencing factories were established to collect thousands of ESTs. As the collections grew, the concept developed that one might compare EST sets produced from related mRNA samples to reveal differentially expressed species. In practice this has not been effective for other than the most highly expressed mRNAs, in part because of the arithmetic of mRNA expression. Most of the mRNA species are expressed in the range between 0.3 and 10 parts in 10^5 ; therefore, for a substantial portion of this abundance class even to be detected, hundreds of thousands of ESTs must be collected. When multiple samples are being compared to judge relative expression levels, the number of ESTs required becomes economically unfeasible. Despite these limitations, EST collections have provided two benefits. First, the snippets of sequence have been a rich source of data for assembling longer mRNA sequences from which putative protein sequences can be discerned. Second, they have changed the scale of RNA expression and DNA sequencing studies, popularizing large-scale descriptive analyses while showing the way to whole genome sequence determination.

A conceptually related but technologically distinct and more systematic EST-like approach is serial analysis of gene expression (SAGE) (Velculescu et al., 1995). In SAGE, short cDNA fragments are produced corresponding to the region (generally 10–15 nucleotides) adjacent to a site for restriction endonuclease cleavage near the 3' ends of the members of an mRNA population. These so-called tag fragments are produced in proportion to the concentration of each mRNA and come from a discrete, defined position; hence, they can be electronically recognized if they come from previously known mRNAs or can be recognized as novel. The tags are incubated with DNA ligase to form long tag polymers, which are cloned and subjected to DNA sequence analysis. Once polymerized, a single sequencing reaction detects tags from dozens of individual mRNAs, thus increasing the

throughput by more than an order of magnitude over EST sampling methods. For cell lines, it is possible to obtain reliable estimations of mRNA concentrations of all but the most rare species if $\geq 250,000$ tags are collected. For complex tissues with many cell types and for more rare mRNAs, many more tags need be collected; thus economic considerations limit how far and to how many RNA samples the analyses are extended, and tend to make SAGE studies monolithic.

Closed-system approaches: gene arrays

As databases accumulated thousands of cDNA and EST sequences, automation and miniaturization technologies were developed to place these sequences in closely packed arrays that could be used as cDNA hybridization targets so as to allow the expression of thousands of genes to be tracked simultaneously (cf. Lockhart et al., 1996). Arrays represent a rapid method for surveying the expression of already identified mRNAs whose concentrations exceed the limit of detection, which is sequence dependent and ranges from one part in 10^4 to one part in 10^5 . As such, they are useful for diagnostic applications for organisms for which a great deal of genomic information has already been accumulated. However, because a sequence must already be in hand before its expression pattern can be queried, they do not represent a gene discovery method per se.

PCR-based methodologies

The advent of PCR and of commercially available, high-throughput thermocycling machines has led to the development of methods for amplification of mRNA populations for electrophoretic display. The initial methods, usually called differential display, used the property of pairs of short (10-mer), arbitrarily chosen oligonucleotides to prime PCRs on complex cDNA mixtures, generating a few dozen amplified products per reaction, although there were several mismatches between each primer and the template cDNAs. The lengths of the products were displayed by electrophoresis. Although the PCR products were mismatched across the primer-binding sites at either of their ends, the rest of their sequences corresponded to portions of mRNAs in the mixture. When such reactions were performed on different cDNA populations and the product peaks examined in adjacent gel lanes, products with different intensities were candidates for portions of differentially expressed mRNAs. By varying the sequences of pairs of such mismatch primers and mixing up the pairs, it was possible to generate thousands of products in a few hundred reactions (cf. Liang and Pardee, 1992). The early PCR-based display methods served as inspirations for the present state-of-the-art approaches, although they themselves, despite some successes, also had shortcomings. Because of the nature of mismatch priming, it was difficult to establish reproducible reaction conditions leading to a high rate of false-positive signals; hence considerable follow-on work was required. The reactions were also biased toward more abundant templates, leading to differential sensitivity. The method also did not lend itself to the informatics revolution that has occurred since large sequence databases and fast computers have emerged.

Open-system PCR-based methods that overcome these limitations have been developed, the most powerful of which is total gene expression analysis (TOGA) (Sutcliffe et al., 2000). In TOGA (Fig. 1), cDNA synthesis is initiated at a fixed point adjacent to the poly(A) tail at the 3' end of each mRNA. The products are treated with a restriction endonuclease recognizing four nucleotides, which cleaves most cDNAs in the proximity of

Figure 1. *A*, The methodological steps in the TOGA cDNA display procedure, adapted from Sutcliffe et al. (2000). *B*, Automation in mRNA expression analysis: here a robot inserts bar-coded 96 well reaction tray into a computer-controlled PCR thermocycling machine. *C*, Comparison of TOGA mRNA expression profiles from eight regions of the mouse brain highlighting six hypothalamus-enriched mRNAs, three of which are already known (vasopressin, hypocretin, and oxytocin) and three are novel (Hy33, Hy88, and Hy94). *D*, A few thousand hypocretin-expressing neurons in the dorsolateral hypothalamus detected by *in situ* hybridization.

their 3' ends. After primer-binding sites are added to either end of the fragments, the fragments are amplified in pools by PCR, using high-fidelity base pairing at the four nucleotides adjacent to the 5' cleavage site (there are 256 primer permutations) to produce 256 nonoverlapping pools of fluorescently labeled products, whose lengths are measured by electrophoresis. These steps assign each mRNA in a sample an address based on its nucleotide sequence: eight contiguous nucleotides (composed of the restriction cleavage site and the four immediately adjacent nucleotides that were used to parse the PCR products into pools), and their distance to the 3' end. One of the advantages of this rather straightforward method was that it was easily amenable to automation on an industrial scale, allowing each of the 256 primers to be individually optimized so that, beginning with modest amounts of mRNA (20 ng), highly reproducible fluorescent product sets are generated and the product lengths are measured and accumulated automatically into a database of mRNA abundance (peak amplitude) and address (eight nucleotides plus length). A single iteration of TOGA on an mRNA sample systematically detects 60–70% of the mRNAs (those that contain a 3' proximal

site for the endonuclease and whose concentrations are above the fluorescence detection limits: approximately one part in 10^6 with present systems), both those previously known and those yet undiscovered. Those mRNAs without a proximal site are collected in subsequent iterations using different restriction endonucleases. The addressing mechanism facilitates computer-rapid assessment of differential mRNA expression patterns while also enabling instantaneous links to nucleotide sequence databases and the literature.

One of the advantages of such an automated format is that it allows several expression criteria to be assessed simultaneously, sifting through literally thousands of mRNAs, a considerable portion of which are presently unknown, to find limited sets that deserve research priority. The utility of this approach is greatly enhanced by the databases, including the human genome drafts, that are the result of the genomics revolution; these databases often allow one to obtain substantial information about the protein-encoding capacity of novel mRNAs at a very early stage of analysis. In a recent application, we used TOGA to measure the accumulation of mRNAs in the mouse striatum after a time

course of chronic exposure to the neuroleptic clozapine (Thomas et al., 2001a). We tracked >11,000 striatal mRNAs and measured substantial increases or decreases in several, including that encoding apolipoprotein D (apoD), suggesting that apoD might be associated with the activity of clozapine in benefiting patients with psychoses. To test this hypothesis, we examined patient material (Thomas et al., 2001b). We measured a significant decrease in the concentration of apoD in serum samples from schizophrenic patients. In contrast, apoD levels were significantly increased in the dorsolateral prefrontal cortex and caudate of schizophrenics. No differences in apoD immunoreactivity were detected in the occipital cortex, hippocampus, substantia nigra, or cerebellum. The low serum concentrations of apoD observed support hypotheses involving systemic insufficiencies in lipid metabolism/signaling in schizophrenia. Elevation of apoD selectively within CNS regions implicated in neuropathology suggests a focal compensatory response that neuroleptic drug regimens may augment.

The neurogenomic millennium

The advent of high-capacity computing and the employment of automation have changed the scale of the investigative process. These advances, and the human genome sequence efforts, have also legitimized descriptive experimental biology, provided it is systematic and thorough, so as to let the nervous system speak for itself in directing its analysis. We anticipate an era during which not only will the major neurological disorders receive mechanistic explanations leading to therapeutic address, but also neural processes that we presently do not even imagine will reveal themselves.

REFERENCES

- Adams MD, Kelley JM, Gocayne JM, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252:1651–1656.
- Barnstable C, Jessell T, Sanes J, Stevens C, Robertson M (1983) How molecular is neurobiology? *Nature* 306:14–16.
- de Lecea L, Criado JR, Prospero-Garcia O, Gautvik KM, Schweitzer P, Danielson PE, Dunlop CLM, Siggins GR, Henriksen SJ, Sutcliffe JG (1996) Cortistatin, a cortical neuropeptide with neuronal depressant and sleep-modulating properties. *Nature* 381:242–245.
- de Lecea L, Kilduff TS, Peyron C, Gao X-B, Foye PE, Danielson PE, Fukuhara C, Battenberg ELF, Gautvik VT, Bartlett FS, Frankel WN, van den Pol AN, Bloom FE, Gautvik KM, Sutcliffe JG (1998) The hypocretins: hypothalamus-specific peptides with neuroexcitatory activity. *Proc Natl Acad Sci USA* 95:322–327.
- Gautvik KM, de Lecea L, Gautvik VT, Danielson PE, Tranque P, Dopazo A, Bloom FE, Sutcliffe JG (1996) Overview of the most prevalent hypothalamus-specific mRNAs, as identified by directional Tag PCR subtraction. *Proc Natl Acad Sci USA* 93:8733–8738.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
- Liang P, Pardee AB (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 257:967–971.
- Lockhart DJ, Dong H, Byrne MC, Follett MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 14:1675–1680.
- Milner RJ, Sutcliffe JG (1983) Gene expression in rat brain. *Nucleic Acids Res* 11:5497–5520.
- Sutcliffe JG, Foye PE, Erlander MG, Hilbush BS, Bodzin LJ, Durham JT, Hasel KW (2000) TOGA: an automated parsing technology for analyzing expression of nearly all genes. *Proc Natl Acad Sci USA* 97:1976–1981.
- Thomas EA, Danielson PE, Nelson PA, Pribyl TM, Hilbush BS, Hasel KW, Sutcliffe JG (2001a) Clozapine increases apolipoprotein D expression in rodent brain: towards a mechanism for neuroleptic pharmacotherapy. *J Neurochem* 76:789–796.
- Thomas EA, Dean B, Pavey G, Sutcliffe JG (2001b) Increased CNS levels of apolipoprotein D in schizophrenic and bipolar subjects: implications for the pathophysiology of psychiatric disorders. *Proc Natl Acad Sci USA* 98:4066–4071.
- Timberlake WE (1980) Developmental gene regulation in *Aspergillus nidulans*. *Dev Biol* 78:497–510.
- Travis GH, Brennan MB, Danielson PE, Kozak CA, Sutcliffe JG (1989) Identification of a photoreceptor-specific mRNA encoded by the gene responsible for retinal degeneration slow (*rd*). *Nature* 338:70–73.
- Usui H, Falk J, Dopazo A, de Lecea L, Erlander MG, Sutcliffe JG (1994) Isolation of clones of rat striatum-specific mRNAs by directional tag PCR subtraction. *J Neurosci* 14:4915–4926.
- Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. *Science* 270:484–487.
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, et al (2001) The sequence of the human genome. *Science* 291:1304–1351.
- Watson JB, Battenberg EF, Wong KK, Bloom FE, Sutcliffe JG (1990) Subtractive cDNA cloning of RC3, a rodent cortex-enriched mRNA encoding a novel 78 residue protein. *J Neurosci Res* 26:397–408.
- Zhuo D, Zhao WD, Wright FA, Yang HY, Wang JP, Sears R, Baer T, Kwon DH, Gordon D, Gibbs S, Dai D, Yang Q, Spitzner J, Krahe R, Stredney D, Stutz A, Yuan B (2001) Assembly, annotation, and integration of UNIGENE clusters into the human genome draft. *Genome Res* 11:904–918.