Jak-STAT Pathways and Transcriptional Activation in Response to IFNs and Other Extracellular Signaling Proteins

James E. Darnell Jr., Ian M. Kerr, George R. Stark

Through the study of transcriptional activation in response to interferon α (IFN-α) and interferon γ (IFN-γ), a previously unrecognized direct signal transduction pathway to the nucleus has been uncovered: IFN-receptor interaction at the cell surface leads to the activation of kinases of the Jak family that then phosphorylate substrate proteins called STATs (signal transducers and activators of transcription). The phosphorylated STAT proteins move to the nucleus, bind specific DNA elements, and direct transcription. Recognition of the molecules involved in the IFN-α and IFN-γ pathway has led to discoveries that a number of STAT family members exist and that other polypeptide ligands also use the Jak-STAT molecules in signal transduction.

Transmembrane receptors enable cells to sense their outside environment. Both polypeptide ligands and ligands of small molecular size bind to specific cell surface receptors with high affinity, causing cells to alter their metabolism in many different ways. There is now ample evidence that one important consequence of ligand-receptor interaction at the cell surface, particularly for receptors that bind extracellular signaling proteins (ESPs), is the transcriptional activation of previously quiescent genes. The extracellular signaling proteins include proteins often referred to as cytokines (for instance, IFN-α, IFN-β, IFN-γ, and the interleukins) or growth factors (for instance, epidermal growth factor, platelet-derived growth factor, growth hormone, and ciliary neurotrophic growth factor). A few genes (such as c-*fos*) are activated immediately by several different, structurally unrelated ESPs (*1–3*), whereas many genes are activated rapidly and specifically only by particular ESPs (*4–10*). Because cells probably have receptors to many dozens of different ESPs simultaneously, a central question arises: How is the specific transcriptional response to a particular ESP achieved?

Over the past decade, experiments exploring the induction of transcription by IFN-α (*5, 7, 8, 11–20*) and IFN-γ (*21–28*) have permitted a general understanding of many individual pathways of signal transduction that connect events at the cell surface directly to gene activation. We refer to these as the Jak-STAT family of pathways; they begin with the binding of ESPs to specific receptors that are associated with protein tyrosine kinases that become activated by ligand attachment. The kinases presently known to function in these pathways belong to the Janus kinase (Jak) family (*29–40*), and their activation is associated with their own phosphorylation on tyrosine residues. The Jak proteins, as we will discuss, can be associated with receptors that either possess or lack tyrosine kinase activities of their own. Latent cytoplasmic proteins termed STATs (signal transducers and activators of transcription) (*36, 41–43*) also become activated, presumably by the Jaks, through phosphorylation of tyrosine residues. The activated STAT proteins are translocated to the nucleus where, by themselves or in combination with otherwise weak DNA binding proteins, they bind to specific sequences (response elements) and stimulate transcription. These conclusions have been reached by two parallel approaches: biochemistry and gene cloning on the one hand (*11–28, 36–43*) and somatic cell genetics on the other (*33–35, 44–47*). The congruent results from these two approaches have provided definitive evidence concerning many of the mechanisms at work in the Jak-STAT class of signal transduction pathways. Our review here first describes the highlights of the research on the IFN pathways and then how this work has helped to suggest roles for the Jak and STAT protein families in response to other polypeptide ligands.

J. E. Darnell Jr. is in the Laboratory of Molecular Cell Biology, Rockefeller University, New York, NY 10021, USA. I. M. Kerr is in the Imperial Cancer Research Fund, Lincoln's Inn Fields, London WC2A 3PX, UK. G. R. Stark is in the Cleveland Clinic Foundation Research Institute, Cleveland, OH 44195, USA.

Genes Transcriptionally Responsive to the IFNs and Their DNA Response Elements

The first step in the study of transcriptional activation of genes by the IFNs was the isolation of complementary DNAs (cDNAs) corresponding to messenger RNAs (mRNAs) that were strongly induced by IFN-α or IFN-γ (*4, 5, 7, 8, 21, 22, 25–28*). Some of these mRNAs are induced only by IFN-α, some by both IFNs, and some only or mainly by IFN-γ. Although both IFN-α and IFN-γ lead to an antiviral response and growth restraint in a number of different cell types, their amino acid sequences are not related and they bind to different receptors (*48*). Because the receptors, the ligands, and the sets of responsive genes are all different, the IFN-α and IFN-γ response pathways were initially thought to be independent.

With the identification of IFN-α–inducible mRNAs, run-on transcriptional analysis showed that transcription of chromosomal genes was enhanced by IFN-α (in some cases by at least 50-fold) within 15 to 30 min and without the need for new protein synthesis (*4, 5, 7, 8, 11*). These results indicated that preexisting proteins were modified in a ligand-dependent manner to effect the increase in transcription. Genomic clones corresponding to the IFN-α–responsive mRNAs were isolated, and the DNA sequences that direct the IFN-α–induced transcriptional response were identified by deletion, site-directed mutagenesis, and transfection analysis (*11, 12, 49–52*). The element responsible for the IFN-α response is a highly conserved region of 12 to 15 base pairs (bp), the ISRE (interferon-stimulated response element) (Table 1). The ISREs placed upstream or downstream of reporter genes activate transcription in an IFN-α–dependent manner, and mutations in the most highly conserved parts of the sequence abrogate the response (*12, 49*). Most genes that respond to IFN-α have ISREs, usually within 200 bp of the transcription start site, which suggests that this sequence (probably influenced by neighboring sequences to some degree) is important for transcriptional activation in response to IFN-α.

Interferon γ stimulates transcription of genes that function in the immune response, such as those encoding class I and class II major histocompatibility antigens, but in several cases only after six or more hours, during which new protein synthesis is required (*53*). Therefore, such genes may not be activated by preexisting proteins. However, IFN-γ also causes immediate transcriptional activation of the gene encoding guanylate binding protein (GBP) (*21–24*) and several other genes (*26–28*), and a consensus immediate response ele-

ment, the IFN-γ activation site (GAS), has been characterized (Table 1). The number of variations of the GAS and surrounding sequences that can participate in immediate transcriptional responses to IFN-γ, IFN-α, or other ESPs is not yet clear, but there is little doubt that a unit of 20 to 30 bases containing different examples of this 9-bp core sequence can serve as the initial response element in the IFN-γ pathway.

ISGF-3 and Characterization of the STAT Proteins

With the development of the simple and specific gel retardation assay using cell extracts and labeled, double-stranded oligodeoxynucleotides (*54, 55*), it was possible to detect and characterize the proteins that bound to the ISREs and to the GAS elements. A number of proteins bind specifically to ISREs (*12, 13, 19, 20*). However, the time course of IFN-α–induced transcription had been determined and the properties of one factor, called interferon-stimulated gene factor–3 (ISGF-3), suggested it as the primary positive regulator. This factor is not present in functional form before IFN-α treatment and appears within minutes after treatment of cells with IFN-α. Its abundance increases and decreases in parallel with changes in the rate of transcription (*4, 12, 14, 56*). Two important properties of ISGF-3 were discovered before the identification and cloning of its constitutent polypeptides. First, the protein or proteins capable of forming ISGF-3 are activated after IFN-α treatment in the cytoplasm (*12–14, 56*). In addition, evidence that ISGF-3 was composed of more than one protein was obtained. Mixing cytoplasmic extracts from IFN-α–treated cells with extracts from HeLa cells pretreated with IFN-γ (*13, 53, 54*) increased formation of ISGF-3. This result indicates that there are two or more components to ISGF-3 and that the pretreatment with IFN-γ that did not by itself induce ISGF-3 formation had prepared the cells to better respond to IFN-α.

Cross-linking of protein in extracts containing ISGF-3 to the ISRE identified a 45- to 50-kD protein that contacted DNA but bound weakly by itself and did not give the characteristic gel-shift pattern (*15, 16*). Partially purified preparations containing ISGF-3 were complexed with the ISRE and subjected to electrophoretic purification. Proteins of 113, 91, and 84 kD and the 48-kD DNA binding protein were recovered (*15*). Partially purified material containing the 113-, 91-, 84- and 48-kD species was prepared, peptide sequences from the four polypeptides were obtained, degenerate oligonucleotides corresponding to the peptides were synthesized, and cDNAs encoding candidate proteins for ISGF-3 were cloned in bacterial plasmids (*17, 18*). Antibodies were prepared to the putative ISGF-3 proteins produced in bacteria. These antibodies bound to the ISGF-3–DNA complex, making it migrate more slowly ("supershifting"), and they detected proteins that were translocated to the nucleus, which indicates that the 113-, 91-, 84-, and the 48-kD proteins do indeed constitute ISGF-3 (*17, 18, 41, 57*).

Table 1. Sequence comparisons of DNA binding sites: ISRE for IFN-α–stimulated genes; GAS for IFN-γ–stimulated genes (*26, 95*). ISG54, IFN-stimulated gene encoding 54 kD; OAS, oligo A synthetase; MIG, monokine induced by IFN-γ; ICSBP, IFN consensus sequence binding protein.

Gene	Nucleotide sequence
IFN-α–stimulated response elements	
ISG54	AGTTTCACTTTCCC
ISG15	AGTTTCGGTTTCCC
6-16	AGTTTCATTTTCCC
9-27	AGTTTCTATTTCCT
OAS	GGTTTCGTTTCCTC
GBP	ACTTTCAGTTTCAT
Consensus	AGTTTCNNTTTCNC/T
IFN-γ–activated site	
GBP	TTACTCTAA
Ly-6E	TTCCTGTAA
FcγR1	TTCCCAGAA
MIG	TTACTATAA
ICSBP	TTCCGAGAA
Consensus	TTNCNNNAA

Sequence analysis of the 48-kD protein showed it to be a member of a known family of DNA binding proteins of previously unknown physiological function (*58–60*). This family includes IRF-1 and IRF-2 (interferon regulatory factors 1 and 2), which in transfection experiments can be shown to be positive and negative acting proteins, respectively, acting through a DNA element (TTCNNTTT) that represents the core but not the entire ISRE (*58–60*). The 48-kD protein also has a weak sequence similarity to the protein encoded by *myb* (*58–60*).

The sequences of the 91- and 84-kD proteins are identical, except that the 38 COOH-terminal amino acids of the 91-kD protein are absent in the 84-kD protein (*18*); we refer to these proteins herein as the 91(84)-kD protein. Thus, differentially spliced mRNA products of a single gene generate the 91(84)-kD protein through alternate polyadenylate addition sites and 3′ exons. The 113-kD protein is about 40% identical to the 91(84)-kD protein over a 700–amino acid region (*18*), which suggests that duplication and evolutionary drift have occurred. The sequences of the 113-kD and 91(84)-kD proteins contain 20– to 30–amino acid segments in the NH_2-terminal halves that might form helices (including one region of heptad leucine repeats) and

Fig. 1. Amino acid sequence in the region of tyrosine phosphorylation for Stat1 and Stat2 (*43, 67*). The boxed Y is phosphorylated. (Sequence numbering is corrected and is available in data banks.) Abbreviations for the amino acid residues are E, Glu; G, Gly; H, His; I, Ile; K, Lys; L, Leu; R, Arg; S, Ser; T, Thr; V, Val; and Y, Tyr.

regions in the COOH-terminal halves that are highly conserved between the genes encoding these proteins and that also match fairly well the Src homology 2 and 3 (SH2 and SH3) domains of other proteins (*17, 18, 61, 62*). The sequences of the 91(84)-kD and the 113-kD proteins do not match those of any previously known proteins and are the first two members of a new protein family. Several other closely related family members have now been found (*63–66*). As discussed below, the 91(84)-kD and the 113-kD proteins serve dual roles as signal transducers as well as activators of transcription; because of this, we suggest the generic name STAT for this family of proteins. To aid in identifying members of this family, we suggest that these proteins be named in the order of their description. Thus, Stat1α and Stat1β are the 91-kD and 84-kD proteins, respectively, Stat2 is the 113-kD protein, and Stat3, Stat4, and Stat5 are newly described members (*63–66*).

Mechanism of STAT protein activation. After treatment of ^{32}P-labeled cells with IFN-α, immunoprecipitation of ^{32}P-labeled proteins with antisera to Stat1 and Stat2 [the 91(84)-kD and 113-kD proteins, respectively] revealed that both Stat1 and Stat2 were phosphorylated on tyrosine residues in response to IFN-α (*41, 42*). (A Stat1 antiserum that precipitates both Stat1α and Stat1β shows Stat1β to be phosphorylated like Stat1α, although most cells tested had much less Stat1β than Stat1α.) A single phosphopeptide was found in each protein, and mapping showed a single phosphorylated residue in each (*43, 67*). This residue is in the same relative position in the two proteins, and the sequences downstream from the site have four of nine amino acids conserved (Fig. 1). The sites do not resemble those around phosphotyrosine residues in other proteins.

Binding factor GAF activated by IFN-γ. After the first GAS element in the gene encoding GBP was delineated, a γ-activated factor (GAF) that binds to this element was detected (*23, 24*). The protein or proteins required to form GAF preexist in the cytoplasm and are activated rapidly in response to IFN-γ without new protein

Table 2. Mutant cell lines defective in response to IFN-α or IFN-γ. Complementation groups U1 to U6 were selected with IFN-α, and groups γ1 and γ2 with IFN-γ. ± indicates that some genes do not respond, whereas others respond well.

Complementation group	Response to ligand			Complementing protein	Reference
	IFN-α	IFN-β	IFN-γ		
U1	−	Partial*	+	Tyk2	(*33*)
U2	±	±	±	p48	(*73*)
U3	−	−	−	Stat1	(*46*)
U4	−	−	−	Jak1	(*35*)
U5	−	−	+	?	(*96*)
U6	−	−	+	Stat2	(*69*)
γ1	+	+	−	Jak2	(*34*)
γ2	+	+	−	?	(*97*)

*U1 mutants retain a small but definite residual response to IFN-β (*44*), which can be increased by selection in IFN-β plus HAT. One clone selected in this way had a substantial response to IFN-β but was still unresponsive to IFN-α (*39*).

synthesis (*23, 42*). Partial purification of GAF revealed a prominent 91-kD protein, and ultraviolet cross-linking of the GAF-GAS complex showed DNA contact with a protein of approximately 90 kD (*42*). GAF activity is inhibited, or the GAF-GAS complex is supershifted, by different antibodies to Stat1α. Stat1α or Stat1β is phosphorylated on tyrosine in response to IFN-γ and translocated to the nucleus (*42*). Again, a single tyrosine containing phosphopeptide was observed. The site of tyrosine phosphorylation on the protein from cells treated with IFN-α or IFN-γ is the same, Tyr^{701} (*36, 43, 67*). (As we will discuss below, only Stat1α, and not Stat1β, activates IFN-γ–dependent transcription.)

After partial purification from ^{35}S-labeled cells, the only labeled protein specifically present in the GAS complex was Stat1 (*43*). The GAF-GAS complex does not react with antibodies to either Stat2 or the 48-kD DNA binding protein. Whether Stat1 alone is the sole activator at all GAS sites is not clear, because activation at some sites correlates with binding to a more complex series of factors than GAF alone (*26, 28*). Although GAF activity is immediately activated in response to IFN-γ as is the transcription of some genes, GAF disappears within 2 to 3 hours, whereas IFN-γ–induced transcription continues much longer (*21, 23, 43*).

Use of Somatic Cell Genetics to Identify Genes Required in IFN Response Pathways

A long-term effort has been under way to identify proteins that participate in the IFN signal-response pathways by isolating mutant cell lines defective in these proteins and then cloning the corresponding cDNAs by complementation (*33–35, 44–47*). Candidate proteins include receptors, receptor-associated molecules, and any intracellular proteins necessary to respond to the signal generated at the membrane. Appropriate mutant cells have been obtained and, in addition to allowing further studies on STAT proteins, have been particularly rewarding in pinpointing the role of the Jak family of protein tyrosine kinases in IFN-dependent signal transduction.

To construct a cell line that allowed identification of mutants in the IFN response, we placed a 1.8-kb upstream region, including the promoter from the IFN-α–inducible gene 6-16 (*19*), upstream of the gene for the bacterial enzyme guanosine phosphoribosyltransferase (gpt) and co-transfected it with the selectable marker hyg^B into human HT-1080 cells lacking hypoxanthine phosphoribosyltransferase ($HPRT^-$) (*44*). The $HPRT^-$ cells cannot grow in hypoxanthine-aminopterin-thymidine (HAT) medium and are resistant to the toxic effects of 6-thioguanine (6-TG). Introduction of IFN-α–regulated *gpt* genes allows recipient $HPRT^-$ cells to grow in HAT plus IFN-α and renders their exposure to 6-TG lethal in the presence of IFN-α. In the absence of IFN, the cells have the same phenotype as the parental $HPRT^-$ cells. A clone of such transfected cells (2fTGH) was obtained and treated with the frameshift mutagen ICR-191, resulting in the isolation of unresponsive mutant cell lines after selection in 6-TG and IFN-α (*44, 45, 47*). Exposure to ICR-191 to about 70% lethality resulted in mutation frequencies from about 10^{-9} to about 10^{-6} after one to five rounds of exposure, respectively. Even the most extreme mutagenesis has allowed recovery of mutant cell lines with phenotypes that can be complemented by single functional cDNAs. Many independent isolates were obtained and classified into complementation groups (Table 2). Selection with IFN-α has yielded six complementation groups (U1 to U6) (*44, 45, 47, 68, 69*).

In summary, mutants with defects in each component of ISGF-3 and in three essential protein tyrosine kinases (Tyk2, Jak1, and Jak2) have been obtained. Complementation groups U3 and U4, both of which proved to be defective in response to either IFN-α or IFN-γ (*35, 45, 47*), provided the first evidence that the two IFN-stimulated pathways might share common proteins. Results obtained with U2 and U3 mutants have corroborated the importance of both the 48-kD DNA binding proteins and Stat1α and Stat1β in IFN-α signaling and have allowed the different roles of Stat84 and Stat91 in IFN-γ signaling to be defined. The U3 cells have no Stat1α or Stat1β protein, which allows the molecular genetics of the Stat1 protein to be explored. In addition, the U1 and U4 mutants (in addition to one mutant isolated by a different strategy) were used to establish the importance of the Jak family of protein tyrosine kinases in the response pathways (*33–35*).

A second strategy was used to isolate mutants specifically defective in response to IFN-γ. The upstream region of an IFN-γ–responsive promoter (taken from a gene named 9-27) was used to regulate expression of the cell surface protein CD2, and fluorescence-activated cell sorting was used to isolate mutant and complemented cells (*34*). Two complementation groups that were unresponsive only to IFN-γ (γ1 and γ2) were identified with this procedure. The selection of mutant cells that fail to induce class II human leukocyte antigen molecules on their surface after treatment with IFN-γ led to the isolation of five additional complementation groups (G1 to G5) in which the major defect appears to be in the secondary class II responses (*68*).

Roles of ISGF-3 Proteins in Mutant U3 and U2 Cells

Complementation of mutants U3A and U2A corroborated that Stat1α or Stat1β and the 48-kD DNA binding protein are indeed components of ISGF-3: U3A cells, which are unresponsive to either IFN-α or IFN-γ, lack the two Stat1 mRNA proteins (*35, 47*). By complementing U3A cells individually with cDNAs encoding Stat1α or Stat1β, the roles of these proteins have been partially discerned. Although Stat1α is better than Stat1β in supporting the IFN-α response of some genes, either of these two proteins restores U3A cells to IFN-α–dependent growth in HAT medium and restores the IFN-α–induced activation of ISGF-3. However, only Stat1α restores U3A cells to IFN-γ responsiveness, even though IFN-γ induces Stat1β phosphorylation on the correct tyrosine residue and the protein is translocated to the nucleus and binds to DNA (*43, 47*). Thus, it appears that the COOH-terminal 38 amino acids present in Stat1α and absent in Stat1β are critical for gene activation through GAS elements. Because in most cells there is

much more Stat1α protein than Stat1β, gene activation would be the expected outcome of Stat1 phosphorylation.

To show that phosphorylation mediates the activity of one of these proteins, Stat91, we changed Tyr^{701}, the single tyrosine residue phosphorylated in Stat91 in response to either IFN-α or IFN-γ, to phenylalanine, and the modified protein was expressed in U3A cells. The mutant protein was not phosphorylated, and no IFN-α or IFN-γ response could be detected in U3 cells complemented with this mutant protein (*43*). Both Stat1α and Stat1β and Stat2 contain sequences similar to the SH2 domains that bind phosphotyrosine in many other proteins (*62, 70*). Conversion of Arg^{602} to Leu, which changes a residue crucial for phosphotyrosine binding in the putative SH2 pocket (*70*), prevented phosphorylation of Stat1α and Stat1β, and this mutant failed to restore responsiveness to IFN-α and IFN-γ (*43*).

In sedimentation rate and gel filtration analyses, Stat1α or Stat1β from untreated cells behaves as a monomer, whereas tyrosine-phosphorylated Stat1α or Stat1β from cells treated with IFN is a dimer (*71*). Inhibiting phosphorylation prevents dimerization and subsequent DNA binding. Mutant U3A cell lines permanently transfected with Stat1β and with an engineered long form of Stat1α (Stat1L) form heterodimers of Stat1β and Stat1L upon treatment with IFN-γ. In vitro, the short and long homodimers are stable but can be induced to dissociate and reassociate by incubation with a low concentration of a Stat1 peptide containing the phosphorylated residue Tyr^{701} ($pTyr^{701}$). Thus, it appears that the Stat1 dimer is formed by reciprocal interaction of the SH2 domain of one subunit with $pTyr^{701}$ of the other subunit in the dimer. This result also raises the possibility that heterodimers between different STAT family members might occur.

Without Stat1α or Stat1β no IFN-α response occurs, but Stat2 protein in IFN-α–treated cells is still phosphorylated to the same extent and on the correct residue. Moreover, the phosphoprotein can enter the nucleus, although with decreased efficiency (*67*). Coprecipitation of Stat1 with Stat2 with an antiserum specific for Stat2 (*41*) shows that their interaction occurs in the cytoplasm after, but not before, IFN-α treatment. Such association of phosphorylated Stat1 and Stat2 has been confirmed by gel exclusion chromatography (*72*).

U2A cells have a truncated and inactive 48-kD protein (*73*), and ISRE-regulated genes in these cells are not responsive to IFN-α until the cells are complemented with the full-length 48-kD protein. The gene encoding IRF-1 (*58*), which is not regulated by an ISRE (*74*) but by a DNA region containing a GAS element, can be activated by IFN-α in U2A cells, as are several other genes, such as the mouse Ly-6E gene (*27*). It is likely that Stat1α, activated by IFN-α treatment, participates in this alternative pathway by binding to sequences in this GAS element of these genes. However, U2A cells do not activate the 1-8 U or 9-27 genes in response to IFN-γ, whereas all other responses to IFN-γ tested in these cells are normal. These genes possibly require an activated factor, distinct from GAF and ISGF-3, in which the 48-kD protein and Stat1 are components.

Complementation of Mutant U1A and the Jak Family in IFN Response Pathways

The mutant cell line 11.1, later renamed U1A (*33*), is unresponsive to IFN-α but responds to IFN-γ normally. Human genomic DNA was used to transfect U1A mutant cells, which were selected in HAT medium for restoration of responsiveness to IFN-α (*33, 44*). Cosmid clones and a cDNA clone were found to restore IFN responsiveness. The cDNA was found to encode Tyk2, a member of a family of large proteins (approximately 1200 amino acids) that possess a sequence related to kinase domains in other tyrosine kinases (*29–32*). Two other members of this family, termed Jak1 and Jak2, have been described. These three proteins are more closely related to each other than to other tyrosine kinases. Each has a tyrosine kinase domain in the COOH-terminal portion of the molecule and a kinase-like domain in the NH_2-terminal half. Sequence comparisons failed to locate any obvious transmembrane domain. The COOH-terminal kinase domain of Jak1, expressed in bacterial cells, leads to tyrosine phosphorylation of a limited number of proteins. The NH_2-terminal kinase-like domain lacks this activity (*30*). Complementation of the U1A cell line with Tyk2 cDNA proved that Tyk2 was required in the IFN-α pathway. Tyk2 itself is phosphorylated on tyrosine in response to IFN-α but not IFN-γ, which fits the phenotype of the U1A mutant (IFN-α^-, IFN-γ^+) (*35, 75*). These results focused attention on the possible role of the Jak family in polypeptide-ligand–induced gene activation.

The mutant cell line γ1A, deficient in response to IFN-γ but normal in response to IFN-α, is complemented by a cDNA encoding Jak2 (*34*). Mutant U4A, defective in response to both IFN-α and IFN-γ, is complemented by a cDNA encoding Jak1 (*35*). Immunochemical and biochemical experiments have given results that are consonant with the genetic data: Treatment of cells with IFN-α caused tyrosine phosphorylation of Tyk2 and Jak1 but not of Jak2, whereas IFN-γ treatment led to phosphorylation of Jak1 and Jak2 but not of Tyk2 (*36, 37*).

Conclusion

At this point, a fairly complete picture of the major intracellular molecules involved in the IFN-α and IFN-γ signal transduction pathways is available and a model of the pathways can be drawn (Fig. 2). An IFN binding subunit of each type of IFN receptor has been cloned and characterized (*76, 77*), and the cloning of a second subunit of the IFN-γ receptor has recently been reported (*78, 79*). In one case, the subunit together with the IFN-γ binding component appears to reconstitute a fully active receptor, whereas in the other only a subset of responses is restored, which suggests that

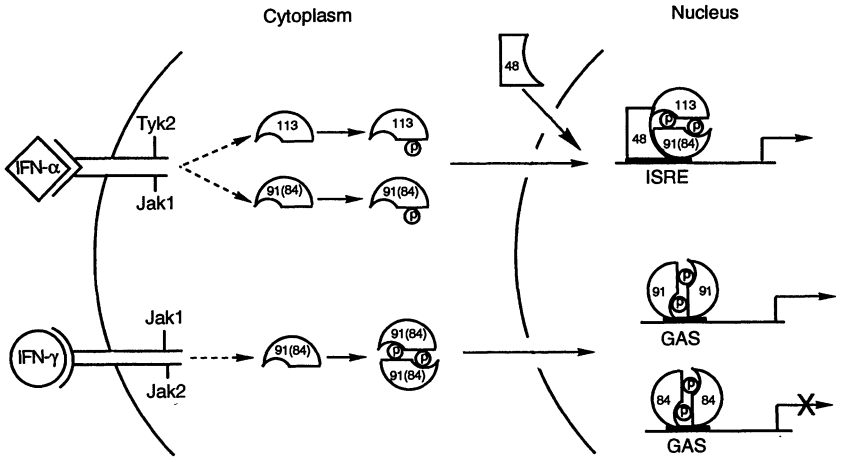

Fig. 2. Diagram of proteins identified in IFN-α– and IFN-γ–dependent signal transduction and gene activation. The Jak kinases are phosphorylated on tyrosine in response to ligand, but the sites and the requirement for such modification are not yet established. The circled P's on the STAT proteins are tyrosine phosphates and the indentations symbolize SH2 domains.

additional or alternative γ receptor subunits may yet be identified. Knowledge of the IFN-α receptor is less advanced, but it also may be complex. For example, immunological evidence for at least three components has been presented (*80*). For simplicity, however, we have indicated only two subunits for each of the receptors in Fig. 2.

As with many growth factors and cytokines, the first interaction in the pathway after ligand receptor interaction may be receptor dimerization. The Jak family members are shown associated with the receptors, and as we will point out later, such an arrangement was first strongly suggested because of the association of Jak family members with other receptors. The evidence for the association of Jak family members with IFN receptors is mainly indirect. U1A cells (lacking Tyk2) or U4A cells (lacking Jak1) do not bind IFN-α well (*33*). Antiserum against Jak1 readily detects this kinase in crude membrane preparations (*81*), and Jak2 co-immunoprecipitates with a subunit of the IFN-γ receptor (*82*). Finally, crude membrane preparations from broken cells can be activated with IFN-α and IFN-γ and produce ISGF-3 or GAF, which implies that there is an association of the kinases with membrane preparations (*83, 84*). In accord with the phosphorylation and genetic data, we assume that Jak1 and Jak2 are associated with the IFN-γ receptor and that Jak1 and Tyk2 are associated with the IFN-α receptor. It remains unclear which kinase functions first. Because Jak1 is activated by both IFN-α and IFN-γ and Stat1 is phosphorylated after treatment with each ligand, perhaps Jak1 is the kinase that directly phosphorylates Stat1 and Tyk2 is the kinase for Stat2. The IFN-γ receptor is reported to be phosphorylated on tyrosine in response to IFN-γ, and Stat1 may bind to this receptor only after phosphorylation (*85*). Whatever the case, a macromolecular assembly including (at least) two chains of the receptor plus two associated kinase molecules is proposed to form a multiprotein complex to which a STAT family member is bound, probably by its SH2 domain (*36*). It is also possible that such complexes could contain one of the many cytoplasmic phosphotyrosine phosphatases that have been described (*86*). We suggest that after IFN-γ treatment and activation of the multiprotein IFN-γ receptor-kinase complex, only Stat1α binds and becomes phosphorylated; dimeric phosphorylated Stat1 then enters the nucleus and, through binding to GAS elements (perhaps in association with other nuclear proteins), activates IFN-γ–dependent transcription. After IFN-α treatment, a receptor complex with phosphorylated Jak1 and Tyk2 allows recognition and phosphorylation of both Stat2 and Stat1. Phosphorylation of Stat1 could lead to transient activation of some genes by both IFN-α and IFN-γ, but because GAF activation is transient the more potent and long-lasting activation of genes by IFN-α requires action through an ISRE. The long-lasting activation of transcription after IFN-γ treatment (*14, 19, 21*) presumably requires some secondary response because GAF activation is transient (*42, 87*).

Jak and STAT proteins in other ESP-activated pathways. Do the Jak and STAT proteins described here or similar proteins function in gene activation after other polypeptides attach to their specific receptors? Many receptors for different cytokines and growth factors (ESPs) are well defined, but in comparison with the IFNs less is known concerning the affected transcription factors or even whether different ligand-specific response elements are used to activate concerted groups of genes. However, recent evidence demonstrates that ligand-receptor interactions other than those with the IFNs can activate the Jak-STAT proteins already known, and several other STAT proteins have now been discovered. Thus, the generality of the Jak-STAT pathways in polypeptide-induced transcription seems likely (*88–93*). For example, the binding site in DNA that controls the transcription of the interleukin-6 (IL-6)–responsive genes and the serum-induced element (SIE) of *c-fos* (especially in its mutated, hyperactive form) can direct activation of plasmid transcription by IL-6 or epidermal growth factor (EGF), respectively, and both these sites resemble the GAS site defined for IFN-γ (*88, 94*). Treatment of cells with IL-6, EGF, or platelet-derived growth factor leads to tyrosine phosphorylation of Stat1 or related proteins and the binding of these proteins to the SIE, GAS, and IL-6 sites (*88–93*). The bandshift complexes induced by IFN-γ, EGF, or IL-6 are not identical, however, and proteins in extracts of different treated cells bind with different affinities to the SIE, GAS, or IL-6 sites. Three bandshift complexes (termed SIFA, SIFB, and SIFC) are produced with the SIE after treatment of epithelial cells with EGF. The slowest of these comigrates with the major IL-6–induced complex from hepatoma cells (*88, 89*). Proteins in the fastest migrating and the middle complexes (SIFC and SIFB, respectively) react with antiserum to Stat1, but those in the slowest migrating band (SIFA) do not. Antibodies to another STAT family member, termed Stat3, show that it is phosphorylated in response to either EGF or IL-6 and that the antibodies react with the middle and top band induced by EGF and the major band induced by IL-6 (*63*). Thus, SIFA and the IL-6–induced complex contain Stat3 and the middle complex, SIFB, contains both Stat1 and Stat3. In addition to Stat3, two other proteins in the family (Stat4 and Stat5) are also known (*64, 65*) that have considerable homology to the other family members. Stat4 has not yet been shown to be phosphorylated in response to a particular ligand. A phosphoprotein activated by prolactin as a DNA binding protein in the sheep casein promoter, the fifth protein in the series (*66*), has about 25% amino acid identity with other family members. These results indicate that the STAT family will provide the cell with a set of proteins that allow a flexible response to different ligands.

A crucial unanswered question is how many clearly distinct response elements exist. Although the IFN-α response element, the ISRE, is clearly different from the IFN-γ element, GAS, the EGF, IL-6, and prolactin elements are all similar to the GAS site. Perhaps the variation around the GAS core sequence will be important in determining specificity for these various factors. The solution to this problem is to find more genes for which transcription is immediately activated by the binding of particular polypeptides and then to characterize the required DNA binding sites and the proteins that bind them.

The finding that proteins other than the IFNs can lead to STAT phosphorylation and activation as DNA binding proteins is paralleled by studies broadening the range of receptors associated with the Jak family. Jak2 or a closely related protein coprecipitates with the erythropoietin receptor, and there is also evidence that Jak2 or a close relative is attached to the growth hormone and IL-3 receptors (*38, 39*). Tyrosine phosphorylation of Jak proteins has been demonstrated in response to such polypeptides as growth hormone, leukemia inhibitory factor, ciliary neurotrophic factor, IL-3, and erythropoietin (*38–40, 88, 89*). Furthermore, other Jak family members may exist, allowing as many different kinase specificities as there are different ligand-receptor-kinase complexes.

Drawing on what we now know about the gene activation pathways triggered by cytokines and growth factors, a generalized scheme for polypeptide-dependent gene control involving Jak and STAT proteins can be considered (Fig. 3). In the pathway, there are four steps where specificity seems to be required for different protein ligands to have different effects on cells.

1) The ligand-receptor interaction is widely acknowledged to be specific.

2) The receptor-kinase complex could also be specific and can obviously involve transmembrane tyrosine kinases like the EGF receptor as well as receptors with no intrinsic kinase activity. Other "soluble" kinases (as the Jaks were formerly regarded) could also be involved. Most if not all

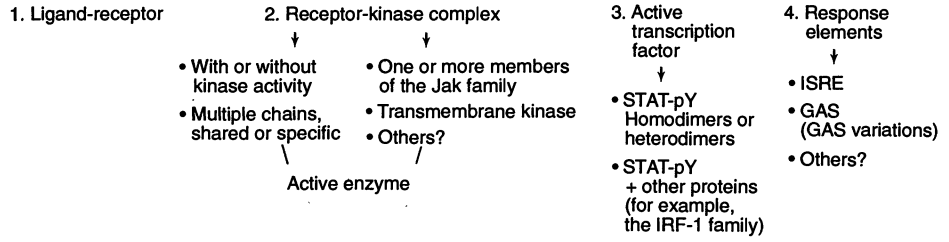

Fig. 3. Points of specificity in the Jak-STAT pathway. 1. The ligand-receptor interaction is accepted as specific. 2. The receptor-kinase complex is likely a multimeric complex, and the active kinase site is unknown. 3. STAT proteins after acting as tyrosine kinase substrates exhibit a variety of mechanisms. 4. Possible variations in DNA binding motifs for the STATs; pY, phosphotyrosine.

cytokine receptors, including the IFN receptors, are composed of two or more chains, some of which can be shared and some of which are specific to a particular ligand. Receptor-associated kinases could have their affinity for substrates altered by the receptor with which they are associated. All these considerations could play a role in determining which kinase finally phosphorylates which STAT family member. It is at present not known in any case which kinase actually phosphorylates any particular STAT protein. For example, IFN-α, IFN-γ, and EGF all lead to the phosphorylation of Jak1 and Stat1, which might suggest a linear pathway with Jak1 as the final active kinase after IFN-α activation of Tyk2 and IFN-γ activation of Jak2. However, in mutants lacking Jak1, neither Tyk2 or Jak2 is phosphorylated after treatment with IFN-α or IFN-γ, which is in contrast to what occurs in wild-type cells. Thus, it does not appear that Tyk2 or Jak2 phosphorylation occurs first, followed by Jak1 phosphorylation. It seems more likely that a functional supercomplex is simply not formed unless both kinases are present to interact with the two (or more) receptor chains. For example, in mutants U1A and U4A that lack Tyk2 and Jak1, respectively, the IFN-α receptor does not bind IFN-α. Thus, the specificities and ranges of activity of the multiprotein receptor-kinase complexes could require all the relevant proteins to interact at once to form a specific complex at the plasma membrane.

3) The available array of STAT proteins and their relative affinities for the receptor-kinase complexes would then determine for each ligand which of the various STAT substrates would become phosphorylated and in what quantities. One or more activated STAT-containing transcription factors would result.

4) The final specificity in any gene activation pathway is nuclear, depending mainly on the variety of DNA elements and the availability of promoter sites that will bind the activated transcription factors. Many details of the IFN pathways are yet to be revealed, and the general pathway proposed for many other extracellular polypeptide ligands remains to be rigorously established. Nevertheless, what has been learned so far provides a picture of how specific ligands can lead to specific gene activation.

REFERENCES

1. J. M. Almendral *et al.*, *Mol. Cell. Biol.* **8**, 2140 (1988).
2. M. Greenberg and E. B. Ziff, *Nature* **311**, 433 (1984).
3. M. E. Greenberg, L. A. Greene, E. B. Ziff, *J. Biol. Chem.* **260**, 14101 (1985).
4. D. Levy and J. E. Darnell Jr., *New Biol.* **2**, 923 (1990).
5. R. L. Friedman, S. P. Manly, M. McMahon, I. M. Kerr, G. R. Stark, *Cell* **38**, 745 (1984).
6. G. H. Murdoch, E. Potter, A. K. Nicolaisen, R. M. Evans, M. G. Rosenfeld, *Nature* **300**, 192 (1982).
7. A. C. Larner *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 6733 (1984).
8. A. C. Larner, A. Chaudhuri, J. E. Darnell Jr., *J. Biol. Chem.* **261**, 453 (1986).
9. T. H. Lee, G. W. Lee, E. B. Ziff, J. Vilcek, *Mol. Cell. Biol.* **10**, 1982 (1990).
10. C. Beadling, K. W. Johnson, K. A. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 2719 (1993).
11. N. Reich *et al.*, *ibid.* **84**, 6394 (1987).
12. D. E. Levy, D. S. Kessler, R. Pine, N. Reich, J. E. Darnell Jr., *Genes Dev.* **2**, 383 (1988).
13. D. E. Levy, D. S. Kessler, R. I. Pine, J. E. Darnell Jr., *ibid.* **3**, 1362 (1989).
14. C. Dale, A. M. A. Iman, I. M. Kerr, G. R. Stark, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 1203 (1989).
15. X.-Y. Fu, D. S. Kessler, S. A. Veals, D. E. Levy, J. E. Darnell Jr., *ibid.* **87**, 8555 (1990).
16. D. S. Kessler, S. A. Veals, X.-Y. Fu, D. E. Levy, *Genes Dev.* **4**, 1753 (1990).
17. C. Schindler, X.-Y. Fu, T. Improta, R. Aebersold, J. E. Darnell Jr., *Proc. Natl. Acad. Sci. U.S.A.* **89**, 7836 (1992).
18. X.-Y. Fu, C. Schindler, T. Improta, R. Aebersold, J. E. Darnell Jr., *ibid.*, p. 7840.
19. A. C. G. Porter *et al.*, *EMBO J.* **7**, 85 (1988).
20. T. C. Dale *et al.*, *ibid.* **8**, 831 (1989).
21. T. Decker, D. J. Lew, Y.-S. Cheng, D. E. Levy, J. E. Darnell Jr., *ibid.*, p. 2009.
22. D. J. Lew, T. Decker, J. E. Darnell Jr., *Mol. Cell. Biol.* **9**, 5404 (1989).
23. T. Decker, D. J. Lew, J. Mirkovitch, J. E. Darnell Jr., *EMBO J.* **10**, 927 (1991).
24. D. Lew, T. Decker, I. Strehlow, J. E. Darnell Jr., *Mol. Cell. Biol.* **11**, 182 (1991).
25. A. D. Luster, J. C. Unkeles, J. V. Ravetch, *Nature* **315**, 672 (1985).
26. R. N. Pearse, R. Feinman, K. Shuai, J. E. Darnell Jr., J. V. Ravetch, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 4314 (1993).
27. K. D. Kahn *et al.*, *ibid.*, p. 6806.
28. Y. Kanno *et al.*, *Mol. Cell. Biol.* **13**, 3951 (1993).
29. A. F. Wilks, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 1603 (1989).
30. ——— *et al.*, *Mol. Cell. Biol.* **11**, 2057 (1991).
31. A. G. Harpur, A. D. Anders, A. Ziemiecki, R. R. Aston, A. F. Wilks, *Oncogene* **7**, 895 (1992).
32. K. Firmbach, I. Byers, T. Shows, R. Dalla-Favera, J. J. Krowlewski, *ibid.* **5**, 1329 (1990).
33. L. Velazquez, M. Fellows, G. R. Stark, S. Pellegrini, *Cell* **70**, 313 (1992).
34. D. Watling *et al.*, *Nature* **366**, 166 (1993).
35. M. Müller *et al.*, *ibid.*, p. 129.
36. K. Shuai *et al.*, *ibid.*, p. 580.
37. O. Silvennoinen, J. N. Ihle, J. Schlessinger, D. E. Levy, *ibid.*, p. 583.
38. B. A. Witthuhn *et al.*, *Cell* **74**, 227 (1993).
39. O. Silvennoinen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 8429 (1993).
40. L. S. Argetsinger *et al.*, *Cell* **74**, 237 (1993).
41. C. Schindler, K. Shuai, V. R. Prezioso, J. E. Darnell Jr., *Science* **257**, 809 (1992).
42. K. Shuai, C. Schindler, V. R. Prezioso, J. E. Darnell Jr., *ibid.* **258**, 1808 (1992).
43. K. Shuai, G. R. Stark, I. M. Kerr, J. E. Darnell Jr., *ibid.* **261**, 1744 (1993).
44. S. Pellegrini, J. John, M. Shearer, I. M. Kerr, G. R. Stark, *Mol. Cell. Biol.* **9**, 4605 (1989).
45. R. McKendry *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 11455 (1991).
46. M. Muller *et al.*, *EMBO J.* **12**, 4221 (1993).
47. J. John *et al.*, *Mol. Cell. Biol.* **11**, 4189 (1991).
48. S. Baron *et al.*, *Interferon: Principles and Medical Applications* (University of Texas Medical Branch at Galveston, Galveston, TX, 1992).
49. D. S. Kessler, D. E. Levy, J. E. Darnell Jr., *Proc. Natl. Acad. Sci. U.S.A.* **85**, 8521 (1988).
50. B. Cohen, D. Peretz, D. Vaiman, D. Beneck, J. Chebath, *EMBO J.* **7**, 1411 (1988).
51. M. N. Rutherford, G. E. Hannigan, B. R. G. Williams, *ibid.*, p. 751.
52. Y. Shirayoshi, R. A. Burke, E. Appella, K. Ozato, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 5884 (1988).
53. M. A. Blanar, E. C. Boettger, R. A. Flavell, *ibid.*, p. 4672.
54. A. Fried and D. M. Crothers, *Nucleic Acids Res.* **9**, 6505 (1981).
55. M. M. Garrer and A. Revzin, *ibid.*, p. 3047.
56. D. E. Levy, D. J. Lew, T. Decker, D. S. Kessler, J. E. Darnell Jr., *EMBO J.* **9**, 1105 (1990).
57. S. A. Veals *et al.*, *Mol. Cell. Biol.* **12**, 3315 (1992).
58. M. Miyamoto *et al.*, *Cell* **54**, 903 (1988).
59. H. Harada *et al.*, *ibid.* **58**, 729 (1989).
60. R. Pine, T. Decker, D. S. Kessler, D. E. Levy, J. E. Darnell Jr., *Mol. Cell. Biol.* **10**, 2448 (1990).
61. X.-Y. Fu, *Cell* **70**, 323 (1992).
62. C. A. Koch, D. Anderson, M. F. Moran, C. Ellis, T. Pawson, *Science* **252**, 668 (1991).
63. Z. Zhong, Z. Wen, J. E. Darnell Jr., *ibid.* **264**, 95 (1994).
64. ———, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
65. K. Yamamato *et al.*, *Mol. Cell. Biol.*, in press.
66. H. Wakao, F. Gouilleux, B. Groner, *EMBO J.*, in press.
67. T. Improta *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
68. C. Mao, D. Davies, I. M. Kerr, G. R. Stark, *ibid.* **90**, 2880 (1993).
69. S. Leung, S. Qureshi, J. E. Darnell Jr., I. M. Kerr, G. R. Stark, unpublished observations.
70. M. Overduin, C. B. Rios, B. J. Mayer, D. Baltimore, D. Cowburn, *Cell* **70**, 697 (1992).
71. K. Shuai *et al.*, *ibid.*, in press.
72. S. Qureshi, M. Salditt-Georgieff, C. Horvath, J. E. Darnell Jr., unpublished observations.
73. M. Muller, D. E. Levy, G. R. Stark, I. M. Kerr, unpublished observations.
74. S. H. Sims *et al.*, *Mol. Cell. Biol.* **13**, 690 (1993).
75. S. Pellegrini and G. R. Stark, personal communication.
76. M. Aguet, Z. Dembic, G. Merlin, *Cell* **55**, 273 (1988).
77. G. Uze, G. Lutfalla, I. Gresser, *ibid.* **60**, 225 (1990).
78. S. Hemmi, R. Böhni, G. Stark, F. D. Marco, M. Aguet, *Cell* **76**, 803 (1994).
79. J. Soh *et al.*, *ibid.*, p. 793.
80. O. R. Colamonici *et al.*, *J. Immunol.* **148**, 2126 (1992).
81. K. Shuai and J. E. Darnell Jr., unpublished observations.
82. D. Guschin, G. R. Stark, I. M. Kerr, unpublished observations.

83. M. David and A. C. Larner, *Science* **257**, 813 (1992).
84. K.-I. Igarashi, M. David, A. C. Larner, D. S. Finbloom, *Mol. Cell. Biol.* **13**, 3984 (1993).
85. R. Schreiber, personal communication.
86. E. H. Fischer, H. Charbonneau, N. K. Tonks, *Science* **253**, 401 (1991).
87. J. Mirkovitch and J. E. Darnell Jr., *Mol. Biol. Cell* **3**, 1085 (1992).
88. H. B. Sadowski, K. Shuai, J. E. Darnell Jr., M. Z. Gilman, *Science* **261**, 1739 (1993).
89. O. Silvennoinen, C. Schindler, J. Schlessinger, D. E. Levy, *ibid.*, p. 1736.
90. S. Ruff-Jamison, K. Chen, S. Cohen, *ibid.*, p. 1733.
91. A. C. Larner *et al.*, *ibid.*, p. 1730.
92. H. Kotanides and N. C. Reich, *ibid.* **262**, 1265 (1993).
93. A. Bonni, D. A. Frank, C. Schindler, M. E. Greenberg, *ibid.*, p. 1575.
94. R. Graham and M. Gilman, *ibid.* **251**, 189 (1991).
95. D. E. Levy, in *Interferon: Principles and Medical Applications*, S. Baron *et al.*, Eds. (University of Texas Medical Branch at Galveston, Galveston, TX, 1992), pp. 161–173.
96. S. Holland, G. R. Stark, I. M. Kerr, unpublished observations.
97. D. Watling, G. R. Stark, I. M. Kerr, unpublished observations.

Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts

Gerard Salton, James Allan, Chris Buckley, Amit Singhal

Vast amounts of text material are now available in machine-readable form for automatic processing. Here, approaches are outlined for manipulating and accessing texts in arbitrary subject areas in accordance with user needs. In particular, methods are given for determining text themes, traversing texts selectively, and extracting summary statements that reflect text content.

Many kinds of texts are currently available in machine-readable form and are amenable to automatic processing. Because the available databases are large and cover many different subject areas, automatic aids must be provided to users interested in accessing the data. It has been suggested that links be placed between related pieces of text, connecting, for example, particular text paragraphs to other paragraphs covering related subject matter. Such a linked text structure, often called hypertext, makes it possible for the reader to start with particular text passages and use the linked structure to find related text elements (*1*). Unfortunately, until now, viable methods for automatically building large hypertext structures and for using such structures in a sophisticated way have not been available. Here we give methods for constructing text relation maps and for using text relations to access and use text databases. In particular, we outline procedures for determining text themes, traversing texts selectively, and extracting summary statements that reflect text content.

Text Analysis and Retrieval: The Smart System

The Smart system is a sophisticated text retrieval tool, developed over the past 30 years, that is based on the vector space model of retrieval (*2*). In the vector space model, all information items—stored texts as well as information queries—are represented by sets, or vectors, of terms. A term is typically a word, a word stem, or a phrase associated with the text under consideration. In principle, the terms might be chosen from a controlled vocabulary list or a thesaurus, but because of the difficulties of constructing such controlled vocabularies for unrestricted topic areas, it is convenient to derive the terms directly from the texts under consideration. Collectively, the terms assigned to a particular text represent text content.

Because the terms are not equally useful for content representation, it is important to introduce a term-weighting system that assigns high weights to terms deemed important and lower weights to the less important terms. A powerful term-weighting system of this kind is the well-known equation $f_t \times 1/f_c$ (term frequency times inverse collection frequency), which favors terms with a high frequency (f_t) in particular documents but with a low frequency overall in the collection (f_c). Such terms distinguish the documents in which they occur from the remaining items.

When all texts or text queries are represented by weighted term vectors of the form $D_i = (d_{i1}, d_{i2}, \ldots, d_{it})$, where d_{ik} is the weight assigned to term k in document D_i, a similarity measure can be computed between pairs of vectors that reflects text similarity. Thus, given document D_i and query Q_j (or sample document D_j), a similarity computation of the form $sim(D_i, Q_j) = \sum_{k=1}^{t} d_{ik}d_{jk}$ can produce a ranked list of documents in decreasing order of similarity with a query (or with a sample document). When ranked retrieval output is provided for the user, it is easy to use relevance feedback procedures to build improved queries on the basis of the relevance of previously retrieved materials.

In the Smart system, the terms used to identify the text items are entities extracted from the document texts after elimination of common words and removal of word suffixes. When the document vocabulary itself forms the basis for text content representation, distinct documents with large overlapping vocabularies may be difficult to distinguish. For example, the vectors covering biographies of John Fitzgerald Kennedy and Anthony M. Kennedy, the current Supreme Court justice, will show many similarities because both Kennedys attended Harvard University, were high officials of the government, and had close relationships with U.S. presidents. The global vector similarity function described earlier cannot cope with ambiguities of this kind by itself. An additional step designed to verify that the matching vocabulary occurs locally in similar contexts must therefore be introduced as part of the retrieval algorithm. This is accomplished by insisting on certain locally matching substructures, such as text sentences or text paragraphs, in addition to the global vector match, before accepting two texts as legitimately similar (*3*).

Consider, as an example, a typical search conducted in the 29-volume Funk and Wagnalls encyclopedia, using as a query the text of article 9667, entitled "William Lloyd Garrison" (Garrison was the best known of the American abolitionists, who opposed slavery in the early part of the 19th century) (*4*). The upper portion of Table 1 shows the top 10 items retrieved in response to a global vector comparison. The top retrieved item is article 9667 itself, with a perfect query similarity of 1.00, followed by additional articles dealing with abolitionism and the slavery issue, retrieved with lower similarity values.

The upper portion of Table 1 consists of relevant items only, with the exception of article 9628, entitled "Gar," retrieved in position eight on the ranked list. Gar is a type of fish, obviously unrelated to the slavery issue but erroneously retrieved because truncated terms were used in the text vectors, and the truncated form of "Garrison" matches "Gar." (Removal of "-ison" as part of the stemming process first reduced "Garrison" to "Garr," as in "comparison" and "compar"; removal of the duplicated consonant then reduced "Garr" to the final

The authors are in the Department of Computer Science, Cornell University, Ithaca, NY 14853–7501, USA.