Serial Analysis of Gene Expression

Victor E. Velculescu, Lin Zhang, Bert Vogelstein, Kenneth W. Kinzler*

The characteristics of an organism are determined by the genes expressed within it. A method was developed, called serial analysis of gene expression (SAGE), that allows the quantitative and simultaneous analysis of a large number of transcripts. To demonstrate this strategy, short diagnostic sequence tags were isolated from pancreas, concatenated, and cloned. Manual sequencing of 1000 tags revealed a gene expression pattern characteristic of pancreatic function. New pancreatic transcripts corresponding to novel tags were identified. SAGE should provide a broadly applicable means for the quantitative cataloging and comparison of expressed genes in a variety of normal, developmental, and disease states.

Determination of the genomic sequence of higher organisms, including humans, is now a real and attainable goal. However, this analysis represents only one level of genetic complexity. The ordered and timely expression of this information represents another level of complexity equally important to the definition and biology of the organism. Techniques based on complementary DNA (cDNA) subtraction or differential display can be quite useful for comparing gene expression differences between two cell types (*1*), but provide only a partial picture, with no direct information about abundance. The expressed sequence tag (EST) approach is a valuable tool for gene discovery (*2*), but like RNA blotting, ribonuclease (RNase) protection, and reverse transcriptase–polymerase chain reaction (RT-PCR) analysis (*3*), it evaluates only a limited number of genes at a time. Here we describe the serial analysis of gene expression (SAGE), a technique that allows a rapid, detailed analysis of thousands of transcripts.

SAGE is based on two principles. First, a short nucleotide sequence tag [9 to 10 base pairs (bp)] contains sufficient information to uniquely identify a transcript, provided it is isolated from a defined position within the transcript. For example, a sequence as short as 9 bp can distinguish 262,144 transcripts (4^9) given a random nucleotide distribution at the tag site, whereas current estimates suggest that even the human genome encodes only about 80,000 transcripts (*4*). Second, concatenation of short sequence tags allows the efficient analysis of transcripts in a serial manner by the sequencing of multiple tags within a single clone. As with serial communication by computers, wherein information is transmitted as a continuous string of data, serial analysis of the sequence tags requires a means to establish the register and boundaries of each tag.

Figure 1 shows how these principles were implemented for the analysis of mRNA expression. Double-stranded cDNA was synthesized from mRNA by means of a biotinylated oligo(dT) primer. The cDNA was then cleaved with a restriction endonuclease (anchoring enzyme) that would be expected to cleave most transcripts at least once. Typically, restriction endonucleases with 4-bp recognition sites were used for this purpose because they cleave every 256 bp (4^4) on average, whereas most transcripts are considerably larger. The most 3′ portion of the cleaved cDNA was then isolated by binding to streptavidin beads. This process provides a unique site on each transcript that corresponds to the restriction site located closest to the polyadenylate [poly(A)] tail. The cDNA was then divided in half and ligated via the anchoring restriction site to one of two linkers containing a type IIS restriction site (tagging enzyme). Type IIS restriction endonucleases cleave at a defined distance up to 20 bp away from their asymmetric recognition sites (*5*). The linkers are designed so that cleavage of the ligation products with the tagging enzyme results in release of the linker with a short piece of the cDNA.

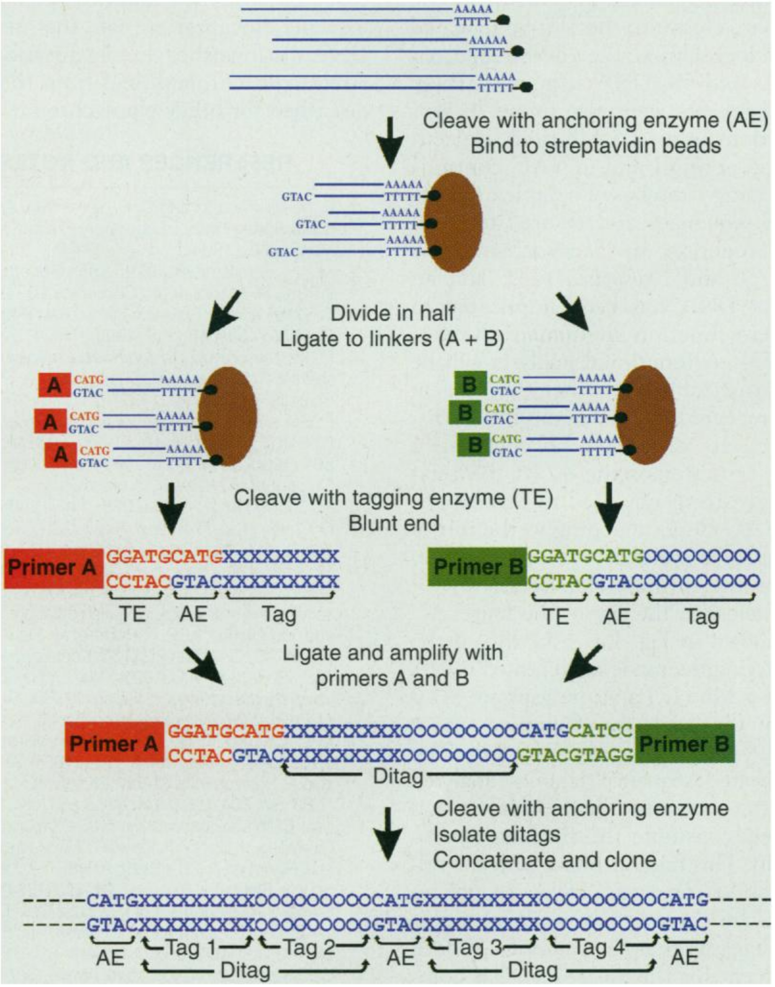
AAAAA
TTTTT
Cleave with anchoring enzyme (AE)
Bind to streptavidin beads
GTAC
Divide in half
Ligate to linkers (A + B)
A
CATG
GTAC
B
Cleave with tagging enzyme (TE)
Blunt end
Primer A
GGATGCATGXXXXXXXXX
CCTACGTACXXXXXXXXX
TE
AE
Tag
Primer B
GGATGCATGOOOOOOOOO
CCTACGTACOOOOOOOOO
Ligate and amplify with
primers A and B
GGATGCATGXXXXXXXXXOOOOOOOOOCATGCATCC
CCTACGTACXXXXXXXXXOOOOOOOOOGTACGTAGG
Ditag
Cleave with anchoring enzyme
Isolate ditags
Concatenate and clone
-----CATGXXXXXXXXXOOOOOOOOOCATGXXXXXXXXXOOOOOOOOOCATG-----
-----GTACXXXXXXXXXOOOOOOOOOGTACXXXXXXXXXOOOOOOOOOGTAC-----
AE
Tag 1
Tag 2
AE
Tag 3
Tag 4
AE
Ditag
Ditag

Fig. 1. Schematic of SAGE. The anchoring enzyme is Nla III and the tagging enzyme is Fok I. Sequences colored red and green represent primer-derived sequences, whereas blue represents transcript-derived sequences, with X and O indicating nucleotides of different tags. See text for further explanation.

V. E. Velculescu and K. W. Kinzler, Oncology Center and the Program in Human Genetics and Molecular Biology, Johns Hopkins University, Baltimore, MD 21231, USA. L. Zhang and B. Vogelstein, Howard Hughes Medical Institute, Johns Hopkins Oncology Center, Baltimore, MD 21231, USA.

*To whom correspondence should be addressed.

For example, Fig. 1 shows a combination of anchoring enzyme and tagging enzyme that would yield a 9-bp tag. After blunt ends were created, the two pools of released tags were ligated to each other. Ligated tags then served as templates for polymerase chain reaction (PCR) amplification with primers specific to each linker. This step served several purposes in addition to allowing amplification of the tag sequences. First, it provided for orientation and punctuation of the tag sequence in a very compact manner. The resulting amplification products contained two tags (one ditag) linked tail to tail, flanked by sites for the anchoring enzyme. In the final sequencing template, this resulted in 4 bp of punctuation per ditag. Second and most importantly, the analysis of ditags, formed before any amplification steps, provided a means to completely eliminate potential distortions introduced by PCR. Because the probability of any two tags being coupled in the same ditag is small, even for abundant transcripts, repeated ditags potentially produced by biased PCR could be excluded from analysis without substantially altering the final results. Cleavage of the PCR product with the anchoring enzyme allowed isolation of ditags that could then be concatenated by ligation, cloned, and sequenced.

As a demonstration of this approach, SAGE was used to characterize gene expression in the human pancreas. We chose Nla III as the anchoring enzyme and Bsm FI as the tagging enzyme, yielding a 9-bp tag (6). Computer analysis of human transcripts from GenBank indicated that greater than 95% of tags of this length were likely to be unique and that inclusion of two additional bases provided little additional resolution (7). As outlined above, mRNA from human pancreas was used to generate ditags (8) that were cloned into a plasmid vector (9). Clones containing at least 10 tags (range 10 to >50) were identified by PCR amplification and manually sequenced (10). Table 1 shows the analysis of the first 1000 tags. Sixteen percent were eliminated because they either had sequence ambiguities or were derived from linker sequences. The remaining 840 tags included 351 tags that occurred once and 77 tags that were identified multiple times (Table 1). Nine of the 10 most abundant tags matched at least one entry in GenBank release 87 (Table 1). The remaining tag was subsequently shown to be derived from amylase (see below). All 10 transcripts were derived from genes of known pancreatic function, and their prevalence was consistent with previous analyses of pancreatic RNA through conventional approaches (11).

The quantitative nature of SAGE was evaluated by construction of an oligo(dT)-primed pancreatic cDNA library that was screened with cDNA probes for trypsinogen 1 and 2, procarboxypeptidase A1, chymotrypsinogen, and elastase IIIB and protease

Table 1. Pancreatic SAGE tags. Tag indicates the 9-bp sequence identifying each tag, adjacent to the 4-bp anchoring Nla III site. *n* and Percent indicate the number of times the tag was identified and its frequency, respectively. Gene indicates the description and accession number of the GenBank release 87 entry found to exactly match the indicated tag when the SAGE software group was used, with the following exceptions. When multiple entries were identified because of duplicated entries (*7*), only one entry is listed. For chymotrypsinogen and trypsinogen 1, other genes (adenosine triphosphatase and myosin alkali light chain, respectively) were identified that were predicted to contain the same tags, but subsequent hybridization and sequence analysis identified the listed genes as the source of the tags. Alu entry indicates a match with a GenBank entry for a transcript that contained at least one copy of the Alu consensus sequence (*15*).

Tag	Gene	*n*	Percent
GAGCACACC	Procarboxypeptidase A1 (X67318)	64	7.6
TTCTGTGTG	Pancreatic trypsinogen 2 (M27602)	46	5.5
GAACACAAA	Chymotrypsinogen (M24400)	37	4.4
TCAGGGTGA	Pancreatic trypsin 1 (M22612)	31	3.7
GCGTGACCA	Elastase IIIB (M18692)	20	2.4
GTGTGTGCT	Protease E (D00306)	16	1.9
TCATTGGCC	Pancreatic lipase (M93285)	16	1.9
CCAGAGAGT	Procarboxypeptidase B (M81057)	14	1.7
TCCTCAAAA	No match (see Table, 2, P1)	14	1.7
AGCCTTGGT	Bile salt stimulated lipase (X54457)	12	1.4
GTGTGCGCT	No match	11	1.3
TGCGAGACC	No match (see Table 2 P2)	9	1.1
GTGAAACCC	21 Alu entries	8	1.0
GGTGACTCT	No match	8	1.0
AAGGTAACA	Secretory trypsin inhibitor (M11949)	6	0.7
TCCCCTGTG	No match	5	0.6
GTGACCACG	No match	5	0.6
CCTGTAATC	M91159, M29366, 11 Alu entries	5	0.6
CACGTTGGA	No match	5	0.6
AGCCCTACA	No match	5	0.6
AGCACCTCC	Elongation factor 2 (Z11692)	5	0.6
ACGCAGGGA	No match (see Table 2, P3)	5	0.6
AATTGAAGA	No match (see Table 2, P4)	5	0.6
TTCTGTGGG	No match	4	0.5
TTCATACAC	No match	4	0.5
GTGGCAGGC	NF-κB (X61499), Alu entry (S94541)	4	0.5
GTAAAACCC	TNF receptor II (M55994), Alu entry (X01448)	4	0.5
GAACACACA	No match	4	0.5
CCTGGGAAG	Pancreatic mucin (J05582)	4	0.5
CCCATCGTC	Mitochondrial CytC oxidase (X15759)	4	0.5
SAGE tags occurring:	Greater than three times	380	45.2
	Three times (15 × 3 =)	45	5.4
	Two times (32 × 2 =)	64	7.6
	One time	351	41.8
	Total SAGE tags	840	100.0

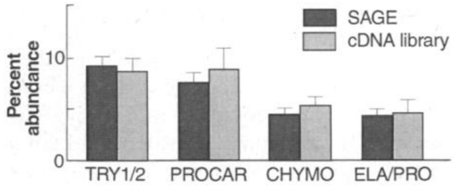

Fig. 2. Comparison of transcript abundance. Bars represent the percent abundance as determined by SAGE (dark bars) or hybridization analysis (light bars). SAGE quantitations were derived from Table 1 as follows: TRY1/2 includes the tags for trypsinogen 1 and 2; PROCAR indicates tags for procarboxypeptidase A1; CHYMO indicates tags for chymotrypsinogen; and ELA/PRO includes the tags for elastase IIIB and protease E. The cDNA hybridizations were as described (*12*). Error bars represent the standard deviation determined by taking the square root of counted events and converting it to a percent abundance. A Poisson distribution was assumed.

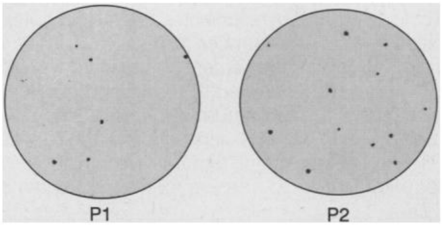

Fig. 3. Screening a cDNA library with SAGE tags. P1 and P2 show typical hybridization results obtained with 13-bp oligonucleotides as described (*13*). P1 and P2 correspond to the transcripts described in Table 2. Images were obtained with a Molecular Dynamics PhosphorImager, and the circle indicates the outline of the filter membrane to which the recombinant phage were transferred before hybridization.

E (*12*). The relative abundance of the SAGE tags for these transcripts was in good agreement with the results obtained with library screening (Fig. 2). Furthermore, whereas neither trypsinogen 1 and 2 nor elastase IIIB and protease E could be distinguished by the cDNA probes used to screen the library (*12*), all four transcripts could readily be distinguished on the basis of their SAGE tags (Table 1).

In addition to providing quantitative information on the abundance of known transcripts, SAGE could be used to identify novel expressed genes. Although for the purposes of SAGE only the 9-bp sequence identifying each transcript was considered, each SAGE tag defines a 13-bp sequence composed of the anchoring enzyme (4-bp) site plus the 9-bp tag. As an illustration of this potential, 13-bp oligonucleotides were used to isolate the transcripts corresponding to four unassigned tags (P1 to P4), that is, tags without corresponding entries from GenBank release 87 (Table 1). In each of the four cases, it was possible to isolate multiple cDNA clones for the tag by simply screening the pancreatic cDNA library with the 13-bp oligonucleotide as hybridization probe (examples in Fig. 3) (*13*). In each case, sequencing of the derived clones identified the correct SAGE tag at the predicted 3′ end of the identified transcript. The abundance of plaques identified by hybridization with the 13-bp oligonucleotides was in good agreement with that predicted by SAGE (Table 2). Tags P1 and P2 were shown to correspond to amylase and preprocarboxypeptidase A2, respectively. No entry for preprocarboxypeptidase A2 and only a truncated entry for amylase was present in GenBank release 87, thus accounting for their unassigned characterization. Tag P3 did not match any genes of known function in GenBank but did match numerous ESTs, providing further evidence that it represented a real transcript. The cDNA identified by P4 showed no significant similarities, suggesting that it represented a previously uncharacterized pancreatic transcript.

These results demonstrate that SAGE can provide both quantitative and qualitative data about gene expression. The combination of different anchoring enzymes with various recognition sites and type IIS enzymes with cleavage sites 5 to 20 bp from their recognition elements lends great flexibility to this strategy. As efforts to fully characterize the genome near completion, SAGE should allow a direct readout of expression in any given cell type or tissue. In the interim, we envision that the major application of SAGE will be the comparison of gene expression patterns in various developmental and disease states. Any laboratory with the capability to perform PCR and manual sequencing could perform SAGE for this purpose. Adaptation of this technique to an automated sequencer would allow the analysis of over 1000 transcripts in a single 3-hour run (*14*).

The appropriate number of tags to be determined will depend on the application. For example, the definition of genes expressed at relatively high levels (0.5% or more) in one tissue, but low in another, would require only a single day. Determination of transcripts expressed at greater than 100 mRNAs per cell (0.025%) should be quantifiable within a few months by a single investigator. Use of different anchoring enzymes will ensure that virtually all transcripts of the desired abundance can be identified. The genes encoding those tags shown to be most interesting on the basis of their differential representation can be positively identified by a combination of database searching, hybridization, and sequence analysis as demonstrated in Tables 1 and 2. Obviously, SAGE could also be applied to the analysis of organisms other than humans and could direct investigation toward genes expressed in specific biologic states.

Table 2. Characterizations of unassigned SAGE tags. Tag and SAGE Abundance are as described in Table 1; 13-mer hyb. indicates the results obtained by screening a cDNA library with a 13-bp oligonucleotide (*13*). The number of positive plaques divided by the total plaques screened is indicated in parenthesis after the percent abundance. A positive in the SAGE Tag column indicates that the expected SAGE tag was identified at the 3′ end of isolated clones. Description indicates the results of BLAST searches of the daily updated GenBank entries at NCBI (National Center for Biotechnology Information) as of 9 June 1995 (*16*). A description and accession number are given for the most significant matches. P1 was found to match a truncated entry for amylase, and P2 was found to match an unpublished entry for preprocarboxypeptidase A2 that was entered after GenBank release 87.

Tag	Abundance (%)		SAGE tag	Description
	SAGE	13-mer hyb.		
P1 TCCTCAAAA	1.7	1.5 (6/388)	+	3′ end of pancreatic amylase (M28443)
P2 TGCGAGACC	1.1	1.2 (43/3700)	+	3′ end of preprocarboxypeptidase A2 (U19977)
P3 ACGCAGGGA	0.6	0.2 (5/2772)	+	EST match (R45808)
P4 AATTGAAGA	0.6	0.4 (6/1587)	+	No match

REFERENCES AND NOTES

1. S. M. Hedrick, D. I. Cohen, E. A. Nielsen, M. M. Davis, *Nature* **308**, 149 (1984); P. Liang and A. B. Pardee, *Science* **257**, 967 (1992).
2. M. D. Adams *et al.*, *Science* **252**, 1651 (1991); M. D. Adams *et al.*, *Nature* **355**, 632 (1992); K. Okubo *et al.* *Nature Genet.* **2**, 173 (1992).
3. J. C. Alwine, D. J. Kemp, G. R. Stark, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5350 (1977); K. Zinn, D. DiMaio, T. Maniatis, *Cell* **34**, 865 (1983); G. Veres, R. A. Gibbs, S. E. Scherer, C. T. Caskey, *Science* **237**, 415 (1987).
4. C. Fields, M. D. Adams, O. White, J. C. Venter, *Nature Genet.* **7**, 345 (1994).
5. W. Szybalski, *Gene* **40**, 169 (1985).
6. Bsm FI was predicted to cleave the complementary strand 14 bp 3′ to the recognition site GGGAC and to yield a 4-bp 5′ overhang (New England BioLabs). Overlapping the Bsm FI and Nla III (CATG) sites as indicated (GGGACATG) would be predicted to result in an 11-bp tag. However, our analysis suggested that under our cleavage conditions (37°C rather than 65°C), Bsm FI often cleaved closer to its recognition site, leaving a minimum of 12 bp 3′ of its recognition site. Therefore, only the 9 bp closest to the anchoring enzyme site was used for analysis of tags.
7. Human sequences (84,300) were extracted from the GenBank release 87 database by means of the FINDSEQ program provided on the IntelliGenetics Bionet on-line service. All further analysis was performed with a SAGE program group written in Microsoft Visual Basic for the Microsoft Windows operating system. The SAGE database analysis program was set to include only sequences noted as "RNA" in the locus description and to exclude entries noted as "EST," resulting in a reduction to 13,241 sequences. Analysis of this subset of sequences with Nla III as anchoring enzyme indicated that 4127 of the 9-bp tags occurred only once, whereas 1511 tags were found in more than one entry. Nucleotide comparison of an arbitrarily chosen subset (100) of the latter entries indicated that at least 83% were a result of redundant database entries for the same gene or highly related genes (>95% identity over at least 250 bp). This suggested that 5381 of the 9-bp tags (95.5%) were particular to a transcript or highly conserved transcript family. Likewise, analysis of the same subset of GenBank with an 11-bp tag resulted in only a 6% decrease in repeated tags (1511 to 1425) instead of the 94% decrease expected if the repeated tags were due to unrelated transcripts.
8. Total pancreatic mRNA (5 μg, Clontech) was converted to double-stranded cDNA with a BRL cDNA synthesis kit. The manufacturer's protocol was used, except for the inclusion of primer biotin-5′-T_{18}-3′. The cDNA was then cleaved with Nla III and the 3′ restriction fragments isolated by binding to magnetic streptavidin beads (Dynal). The bound DNA was divided into two pools, and one of the following linkers was ligated to each pool: linker A, 5′-TTTTACCAGCTTATTCAATTCGGTCCTCTCGCACAGGGACATG-3′, 3′-dideoxyATGGTCGAATAAGTTAAGCCAGGAGAGCGTGTCCCT-5′; linker B, 5′-TTTTTGTAGACATTCTAGTATCTCGTCAAGTCGGAAGGGACATG-3′, 3′-dideoxyAACATCTGTAAGATCATAGAGCAGTTCAGCCTTCCCT-5′. After extensive washing to remove unligated linkers, the linkers and adjacent tags were released by cleavage with Bsm FI. The resulting overhangs were filled in with T4 polymerase, and the pools were combined and ligated to each other. The desired ligation product was then amplified for 25 cycles with 5′-CCAGCTTATTCAATTCGGTCC-3′ and 5′-GTAGACATTCTAGTATCTCGT-3′ as primers. The PCR reaction was then analyzed by polyacrylamide gel electrophoresis (PAGE) and the desired product excised. An additional 15 cycles of PCR were then performed to generate sufficient product for efficient ligation and cloning.
9. The PCR product was cleaved with Nla III and the band containing the ditags was excised and self-ligated. After ligation, the concatenated ditags were separated by PAGE and products greater than 200 bp were excised. These products were

cloned into the Sph I site of pSL301 (Invitrogen). Colonies were screened for inserts by PCR with T7 and T3 sequences located outside the cloning site as primers.

10. Selected clones were manually sequenced as described [G. Del Sal, G. Manfioletti, C. Schneider, *Biotechniques* **7**, 514 (1989)] with 5′-GACGTCGACCTGAGGTAATTATAACC-3′ as primer. Sequence files were analyzed by means of the SAGE software group (*7*), which identifies the anchoring enzyme site with the proper spacing and extracts the two intervening tags and records them in a database. The 1000 tags were derived from 413 unique ditags and 87 repeated ditags. The latter were counted only once to eliminate potential PCR bias of the quantitation, as described in the text.
11. J. H. Han, L. Rall, W. J. Rutter, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 110 (1986); J. Takeda, H. Yano, S. Eng, Y. Zeng, G. I. Bell, *Hum. Mol. Genet.* **2**, 1793 (1993).
12. Pancreatic mRNA from the same preparation was used for SAGE and to construct a cDNA library in the ZAP Express vector. The ZAP Express cDNA Synthesis kit (Stratagene) was used according to the manufacturer's protocol. Analysis of 15 randomly selected clones indicated that 100% contained cDNA inserts. Plates containing 250 to 500 plaques were hybridized as described [J. M. Ruppert *et al.*, *Mol. Cell. Biol.* **8**, 3104 (1988)]. The cDNA probes for trypsinogen 1, trypsinogen 2, procarboxypeptidase A1, chymotrypsinogen, and elastase IIIB were derived by RT-PCR from pancreas RNA. The sequences of primers are available from the authors upon request. The trypsinogen 1 and 2 probes were 93% identical and hybridized to the same plaques under the conditions used. Likewise, the elastase IIIB probe was >95% identical to protease E.
13. Plates containing 250 to 2000 plaques were hybridized to oligonucleotide probes with the same conditions previously described for standard probes except that the hybridization temperature was reduced to room temperature (*12*). Washes were performed in 6× standard saline citrate–0.1% SDS for 30 min at room temperature. The probes consisted of 13-bp oligonucleotides that were labeled with [γ^{32}-P]ATP through use of T4 polynucleotide kinase.
14. An ABI 377 sequencer can produce a 451-bp read for 36 templates in a 3-hour run [(451 bp/11 bp per tag) × 36 = 1476 tags].
15. P. L. Deininger *et al.*, *J. Mol. Biol.* **151**, 17 (1981).
16. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lippman, *J. Mol. Biol.* **215**, 403 (1990).

17. Supported by NIH grants CA57345, CA35494, GM07309. B.V. is an American Cancer Society Research Professor and an Investigator of the Howard Hughes Medical Institute. We thank S. Kern, B. D. Nelkin, and members of our laboratories for their critical review and valuable discussions.

14 June 1995; accepted 11 August 1995

TECHNICAL COMMENT

The Radius of Gyration of an Apomyoglobin Folding Intermediate

Apomyoglobin (apoMb) forms a stable compact partially folded state under acidic conditions (*1*). This "molten globule" intermediate is slightly expanded relative to the native form of the protein, with a radius of gyration (R_g) of 23 (± 2) Å versus 19 (± 1) Å (*2*), and shows stable secondary structure (*3*) in the A, G, and H helices (Fig. 1).

We demonstrated recently, with the use of stopped-flow circular dichroism and pulse-labeling hydrogen exchange measurements, that the earliest detectable intermediate (formed within 6 ms) in the apoMb kinetic refolding pathway closely resembles the equilibrium molten globule state populated under acid conditions (*4*). A key question remained as to how compact this kinetic intermediate is compared to the equilibrium and native states. The cooperative unfolding of the kinetic intermediate and the significant protection from amide proton exchange (as compared to corresponding isolated peptides in solution) led us to propose that the kinetic intermediate is also compact (*4, 5*). Such a proposal could best be verified by direct determination of the size of the protein as it folds, but measurements of this nature were not feasible at the time.

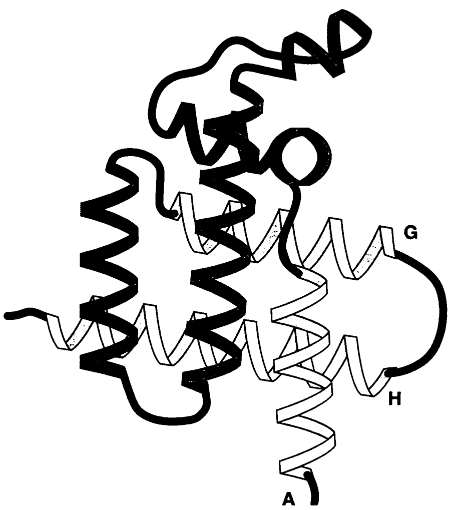

Fig. 1. Sketch of the structure of holo-myoglobin, illustrating the location of the A, G, and H helices, which are present in both the equilibrium and kinetic folding intermediates of the apoprotein.

Newly developed improvements in time-resolved small angle x-ray scattering (SAXS) experiments allow direct measurement of the time-dependent change of R_g of a protein as it folds in the millisecond to second time frame (*6, 7*). We initiated studies of the refolding of apoMb using this technique, under conditions similar to those employed in our previous work (*4*). SAXS data collected during the first 100 ms after initiation of the refolding reaction (*8*) are shown in Fig. 2.

Data collected from the fully refolded protein and unfolded protein are given for comparison (Fig. 2). The data obtained 100 ms after the initiation of folding are within experimental error of the data obtained for the refolded protein, and easily distinguishable from data obtained for the unfolded state. An R_g value of 23 (± 2) Å is obtained at 100 ms, only 1 Å greater than the 22 (± 1) Å value obtained for the refolded protein. By contrast, the unfolded state has an R_g of 34 (± 2) Å. The slightly higher than expected R_g value obtained for the refolded state may result from either experimental error (*9*) or a small degree of sample aggregation owing to radiation damage during exposure. It is possible that the R_g value obtained at 100 ms is similarly inflated, and it may therefore be considered an upper bound on the true R_g.

Our conclusion that the intermediate is compact is based on the small differences

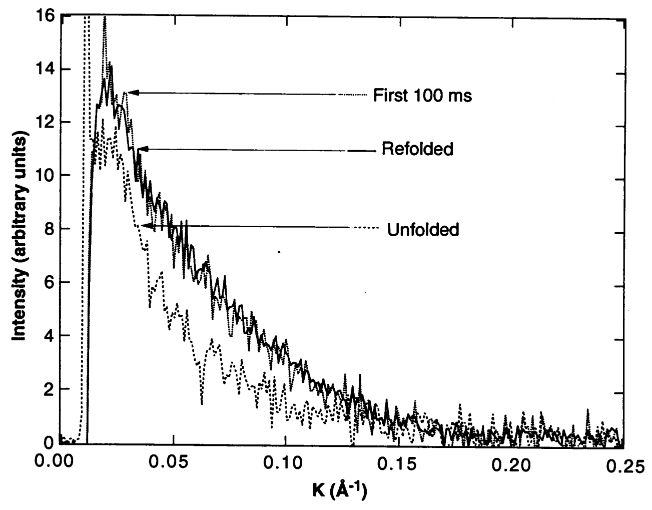

Fig. 2. SAXS data from sperm whale apomyoglobin after 100 ms of folding, after 4.2 s of folding, and in the unfolded state. Detected intensity is plotted as a function of *K*. Data from the unfolded state is scaled to match the folded state data at zero scattering angle. The data obtained from the fully folded protein and that obtained after 100 ms of folding are barely distinguishable from each other and are different from the data for the unfolded protein.