Vol. 141, No. 1, 1986 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
November 26, 1986 Pages 326-333

# CHARACTERIZATION OF SITES OF TYROSINE SULFATION IN PROTEINS AND CRITERIA FOR PREDICTING THEIR OCCURRENCE

Glen Hortin, Rodney Folz, Jeffrey I. Gordon, and Arnold W. Strauss

Department of Biological Chemistry, Washington University, St. Louis, MO

Received October 8, 1986

SUMMARY: A wide variety of secretory proteins have recently been found to undergo post-translational sulfation of specific tyrosine residues. Here, amino acid sequences surrounding known sulfation sites in proteins are analyzed in order to identify factors which determine the specificity of sulfation. Several distinctive features of sulfation sites are identified, including: 1) abundance of acidic amino acid residues, 2) lack of basic residues, 3) low hydropathy, 4) absence of neighboring cysteine residues, 5) lack of extended secondary structure. Rules are proposed for predicting likely sites of sulfation based on the amino acid sequence of a protein. 

Recent studies by Huttner and coworkers (1-3) indicate that sulfation of tyrosine residues is a common modification of secretory proteins. This post-translational modification occurs in a broad phylogenetic range of organisms, and a number of proteins in each tissue examined contain tyrosine sulfate (2, 3). Frequent occurrence of tyrosine sulfate was not recognized previously simply because this modified amino acid is acid-labile. Tyrosine sulfate is degraded by standard methods for compositional and sequence analysis of proteins (2, 4) so that its presence in a protein would not be detected, even if the protein has been sequenced in its entirety by standard techniques. Recognition of the widespread occurrence of the sulfation of proteins has stimulated efforts to identify proteins that contain this modification. Examples of human proteins recently found to contain tyrosine sulfate include: the fourth component of complement (4), fibronectin (5), α-fetoprotein (5), type III procollagen (6), $\alpha_2$-antiplasmin (7, 8), and heparin cofactor II (9). The sulfation of tyrosine residues in proteins appears to be a highly selective process. Only specific tyrosine residues are modified in proteins that are substrates for sulfation. However, the structural determinants of the site specificity are not known. The present study analyzes amino acid sequences adjoining known sites of sulfation in order to clarify how specific tyrosine residues are recognized for sulfation.

0006-291X/86 $1.50

## MATERIALS AND METHODS

Protein database: Amino acid sequences adjacent to 15 known sites of sulfation in 10 proteins were chosen as a database (Table I). The sources of amino acid sequences and identification of sites of sulfation are indicated. Sulfation sites have been determined in B-fibrinopeptides from a number of animal species (10), but, due to extensive homology of these sequences, only one was included in the database. Also, an assumption was made that small peptides known to contain tyrosine sulfate undergo sulfation while the peptides are still segments of larger precursor polypeptides. Validity of this assumption is supported by the observation that the single-chain precursor to C4 (4) and high-molecular weight forms of gastrin contain tyrosine sulfate (11).

Computer analysis: The computer program PARA-SITE[1] analyzes structural parameters as a function of distance from a specific index position (position 0). In this study, tyrosine sulfate residues are used as the index position. PARA-SITE calculates and plots the mean structural parameter value and the standard deviation for all amino acids at each distance from the index position. Secondary structure values used to describe α-helix, β-sheet, and β-turns were obtained from Chou and Fasman (12). Hydropathy values were from Kyte and Doolittle (13). PARA-SITE was written in the C programming language (14). All programs used in this report were run on a VAX 11/785 or MicroVAX II running VMS 4.4 (Digital Equipment Corporation).

## RESULTS AND DISCUSSION

The amino acid sequences listed in Table I comprise the database used for analysis of sites of sulfation. No consensus sequence for sulfation sites is identified among these amino acid sequences. However, several distinctive features of these sequences surrounding sites of sulfation are evident. Most striking is the unusual distribution of several amino acids around sites of sulfation. The distribution of amino acids at different distances from tyrosine sulfate residues (designated as position 0) is tabulated in Table II. There is a very high concentration of acidic amino acid residues in the segment extending 5 residues either side of tyrosine sulfate residues. Within this segment, 46% of the residues are aspartic or glutamic acid, and each example in the database contains at least 3 acidic residues. The strongest preference for acidic residues is at the -2 and -1 positions, where 73% of all amino acids are aspartic or glutamic acid. Each example in the database contains at least one acidic residue at these two positions. In contrast, there are few basic residues in the -5 to +5 region. Only 3% of the total consists of basic residues, and no site of sulfation has more than one basic residue in this segment. Beyond the limits of the -5 to +5 segment there is less preference for acidic residues over basic residues. All 20 amino acids, except for cysteine, occur within the -5 to +5 segment. The closest cysteine residue to any of the

[1] Further details of this program and its general applicability for analyzing protein sequences will be described elsewhere (Folz, R.J. and Gordon, J.I., manuscript in preparation).

TABLE I

SEQUENCES ADJACENT TO TYROSINE SULFATE RESIDUES

| | |
|---|---|
| 1) | -LQIEVTVKGHVEYTMEANEDYEDYEYDELPAKDDPDAPLQPVTPLQ- |
| 2) | -PRGDKLFGPDLKLVPPMEEDYPQFGSPK |
| 3) | -FHKENTVTNDWIPEGEEDDDYLDLEKIFSEDDDYIDIVDSLSVSPTDSDVSAGN- |
| 4) | -SRRWAIHTSEDALDASELEHYDPADLSPTESSLDLLGLNRT- |
| 5) | -RKLVQAYQQRYNLQPYETTDYSNEEQSQRSSSEEQQTQRRK- |
| 6) | -EGTPKPQSHNDGDFEEIPEEYLQ |
| 7) | pQFPTDYDEGQDDRPKVGLGARGHRPY |
| 8) | -HLVADPSKKQGPWLEEEEEAYGWMDFGRRSAEDEN |
| 9) | -GAPQQREANDERRFADGQQDYTGWMDFGRRDDEDDVNERDV- |
| 10) | -VSMIKNLQSLDPSHRISDRDYMGWMDFGRRSAEEYEYTS |

Amino acid sequences extending 20 residues in both directions from sites of sulfation are presented using the single letter code. Sulfated tyrosine residues are underlined. The symbol pQ represents a pyroglutamyl residue. Sources of data on the amino acid sequences of sites of sulfation are shown below:
1) Fourth component of complement (C4), Human (4, 17, 18)
2) Alpha-2-Antiplasmin, Human (7, 8)
3) Heparin cofactor II, Human (9, 19)
4) Coagulation factor X, Bovine (20)
5) Yolk protein 2, Drosophila melanogaster (16, 21)
6) Hirudin, Leech (Hirudo medicinalis)(22)
7) Fibrinogen, Bβ-chain, Bovine (10, 23)
8) Progastrin, Human (24, 25)
9) Procaerulein IV, Frog (Xenopus laevis), (contains up to 4 copies of caerulein)(26)
10) Procholecystokinin, Porcine (27, 28)

sites of sulfation is located more than 20 residues away. Possible significance of the striking absence of cysteine residues is discussed below.

The hydropathy and secondary structure of sites of sulfation were analyzed by computational techniques described in Methods. As expected, considering the abundance of highly polar acidic amino acids adjacent to sites of sulfation, these sites have a very low average hydropathy value. This is especially pronounced in the segment immediately preceding the sulfated tyrosine residue. This segment appears as a prominent valley in a plot of hydropathy values (Fig. 1). The low hydropathy of sites of sulfation may be an important characteristic, because sulfation is a post-translational modification of proteins. Sites will be accessible to the action of a sulfotransferase only if they are exposed on the surface of a protein. The analysis of the distribution of amino acid residues suggests that sites of sulfation are not selected simply on the basis of low hydropathy, however. If that were the major factor involved in the selection of sites of sulfation, a high abundance of

TABLE 2
DISTRIBUTION OF AMINO ACIDS AROUND SITES OF SULFATION

| Amino Acid | Position Relative to Site of Sulfation | | | | | | | | | | | | | | | | | | | | | | | | | Sum | Sum (-5 to +5) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | -12 | | -10 | | -8 | | -6 | | -4 | | -2 | | 0 | 2 | | 4 | | 6 | | 8 | | 10 | | 12 | | | |
| Asp | 0 | 5 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 4 | 2 | 9 | | 3 | 3 | 1 | 2 | 4 | 2 | 1 | 1 | 1 | 4 | 2 | 4 | 57 | 30 |
| Glu | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 5 | 7 | 3 | 7 | 4 | | 3 | 2 | 1 | 4 | 0 | 0 | 1 | 0 | 2 | 0 | 1 | 3 | 54 | 36 |
| TyrS04 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 2 | 0 | 15 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23 | 23 |
| Lys | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 9 | 1 |
| Arg | 0 | 1 | 0 | 1 | 2 | 1 | 2 | 1 | 0 | 0 | 1 | 0 | | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 4 | 2 | 0 | 0 | 0 | 17 | 2 |
| His | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 1 |
| Ala | 0 | 0 | 0 | 2 | 0 | 1 | 2 | 1 | 1 | 1 | 0 | 1 | | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 2 | 2 | 0 | 16 | 6 |
| Asn | 2 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 2 |
| Cys | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Gln | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | | 0 | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 11 | 7 |
| Gly | 2 | 0 | 1 | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | | 1 | 2 | 1 | 1 | 0 | 1 | 2 | 0 | 0 | 0 | 1 | 0 | 17 | 6 |
| Ile | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 4 |
| Leu | 1 | 1 | 2 | 0 | 3 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | | 2 | 0 | 2 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 19 | 7 |
| Met | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | | 1 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 5 |
| Phe | 0 | 0 | 1 | 0 | 1 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | | 0 | 0 | 1 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 11 | 3 |
| Pro | 1 | 0 | 0 | 2 | 1 | 0 | 2 | 1 | 0 | 2 | 0 | 0 | | 1 | 1 | 0 | 1 | 0 | 2 | 1 | 1 | 2 | 0 | 2 | 1 | 21 | 6 |
| Ser | 2 | 0 | 0 | 0 | 1 | 0 | 1 | 2 | 2 | 0 | 0 | 0 | | 1 | 1 | 0 | 1 | 1 | 3 | 0 | 2 | 2 | 4 | 2 | 0 | 25 | 8 |
| Thr | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 2 | 0 | | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 11 | 6 |
| Trp | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 3 |
| Tyr | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 1 |
| Val | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 5 | 1 |
| Acidic Residues | 1 | 6 | 5 | 3 | 2 | 3 | 3 | 7 | 8 | 11 | 0 | 13 | | 6 | 7 | 3 | 6 | 5 | 2 | 3 | 1 | 3 | 4 | 3 | 7 | 134 | 89 |
| Basic Residues | 2 | 2 | 0 | 2 | 3 | 2 | 2 | 1 | 0 | 0 | 1 | 1 | | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 5 | 3 | 0 | 1 | 0 | 30 | 4 |

basic and other polar residues as well as acidic residues would occur near sulfation sites. Furthermore, studies using synthetic peptides as the substrates for protein tyrosine sulfotransferase indicate that acidic residues increase the affinity of peptides for the enzyme (15).

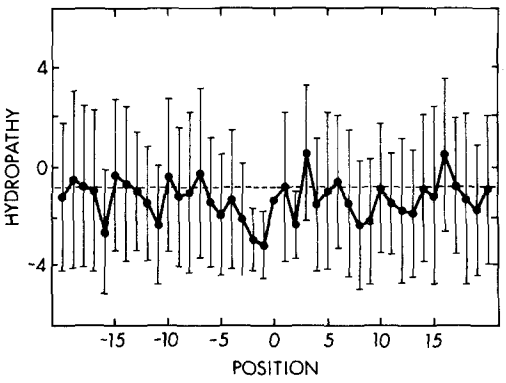


Fig. 1. Plot of average hydropathy values around sites of sulfation. Average hydropathy values at each postition relative to sites of sulfation (Position 0) were calculated for the amino acid sequences shown in Table 1. Error bars indicate the standard deviation of all values at a given position. The reference line across the middle of the diagram corresponds to the mean hydropathy value of the 20 different amino acids which occur naturally in proteins.

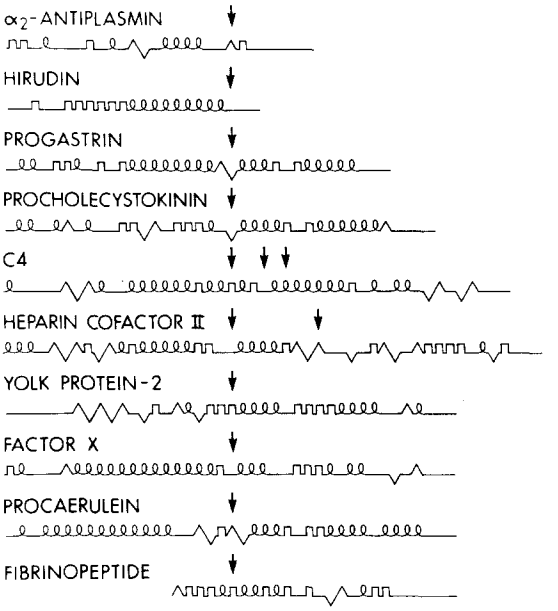


Fig. 2. Predicted secondary structure around sites of sulfation. Secondary structures of the amino acid sequences in Table 1 were determined using the method of Chou and Fasman (12). Arrows indicate sites of sulfation. Each symbol represents a single amino acid residue. Residues occurring in α-helices are shown as coils, β-sheets as zig-zags, β-turns as rectangular steps, and residues without predicted secondary structure are indicated as a horizontal line.

The secondary structure around each site of sulfation in the database was analyzed using the method of Chou and Fasman (12)(Fig. 2). This approach predicted that, many of the proteins will contain α-helical segment immediately before and after sulfation sites. However, tyrosine residues have a low propensity to form α-helices. As a consequence, some of the sites of sulfation are predicted to be located within short gaps between two helical segments. Only one sulfation site (the third site in C4) was predicted to lie within a helical segment, although 42% of the residues in the database were predicted to be in α-helices. That one example was the only case in which a site of sulfation was predicted to be within an extended segment (greater than three residues) of α-helix or β-sheet. Tyrosine residues favor the formation of β-sheet, but there was very low propensity to form extended β-sheets encompassing sites of sulfation. All but three of the sites of sulfation were closely preceded by α-helical segments. In two of the exceptions, yolk protein-2 and the first sulfation site in procholecystokinin, helical segments closely followed the sulfation site. The general conclusions suggested by predictions of secondary structure are that sulfation sites occur within short segments which lack extended secondary structure or are part of β-turns and that sulfation sites are closely flanked by α-helices. Similar conclusions

were reached by using the Para-Site program to average all amino acid sequences in the database and to generate single values at each position for the propensities to form α-helix, β-sheet, and β-turn conformations. Occurrence of sulfation sites at β-turns or at positions lacking secondary structure may contribute to accessibility of these sites. The significance of the probable occurrence of α-helices adjacent to most sulfation sites is not clear.

A number of factors may contribute to maximal exposure of sulfation sites at the surface of proteins. As already noted, known sulfation sites are bounded by multiple acidic amino acid residues, have low hydropathy, and lack extended secondary structure. In addition, a considerable proportion of sulfation sites, 7 out of the 15 in the database, occur within 20 residues of the termini of proteins. Furthermore, there is a complete absence of cysteine residues near sulfation sites. This may be important in permitting optimal access to segments that contain sites of sulfation. Formation of disulfide bonds with other segments of the polypeptide chain would introduce steric hindrance and restrict the flexibility of sites of sulfation. Extrapolation of the principle that accessibility is one of the most important characteristics of sites of sulfation suggests that sulfation sites may occur in extended segments of proteins in preference to globular domains.

Based on the foregoing analysis of amino acid sequences surrounding sulfation sites, five simple rules were empirically derived to aid in predicting the location of sites of sulfation. Tyrosine residues that are likely sites of sulfation are identified by the following criteria:

1) There is an acidic residue at position -1 or -2.
2) There are at least 3 acidic amino acid residues within 5 residues (positions -5 to +5) of the tyrosine residue.
3) No more than 1 basic amino acid residue are within 5 residues of the tyrosine.
4) No more than 3 hydrophobic residues (Ile, Leu, Phe, and Val) are within 5 residues of the tyrosine.
5) No cysteine residues are within 15 residues of the tyrosine.

Most sites of sulfation are expected to conform to all five of the criteria above. There are three known examples, canine fibrinopeptide B, bovine gastrin, and feline gastrin, in which sulfation sites conform with only four of the criteria (10). In those three cases, acidic residues are absent from the -1 and -2 positions. Application of these rules correctly identifies the 3 tyrosine sulfate residues in the fourth component of complement (4) and the 2 in heparin cofactor II (9) and excludes the other 59 tyrosine residues in these two proteins. Moreover, the rules correctly predict the absence of tyrosine sulfate in human serum albumin, which contains 18 tyrosine residues. Two sites of sulfation in yolk protein

2 are predicted by the above rules, but only one site was found experimentally (16). Some sites identified as potential sites of sulfation may not be sulfated due to steric hindrance of sites by higher order structure of the protein or by oligosaccharide chains. This would parallel the situation with N-linked glycosylation, in which not all potential sites identified by amino acid sequence are glycosylated. It is unlikely that any predictive method based solely on the amino acid sequence of proteins will perfectly predict sites of sulfation. Nevertheless, these rules should aid considerably in identifying sites of sulfation in proteins known to contain tyrosine sulfate and in predicting which proteins contain tyrosine sulfate. Identification of sites of sulfation may be essential for complete understanding of the structure and function of many proteins. The effect of sulfation on the biological activity of proteins has not been determined, but sulfation is known to have a profound effect on the activity of some peptides such as cholecystokinin (Refs. in 4).

Acknowledgements: We thank Mark Boguski for reviewing the manuscript. These studies were supported by grants from the NIH and from Monsanto Corp. R.J.F. is supported by a Medical Scientist Training Program Grant (NIH GM-07200).

## REFERENCES

1. Huttner, W. B. (1982) Nature 299, 273-276.
2. Huttner, W. B. (1984) Methods Enzymol. 107, 200-223.
3. Hille, A., Rosa, P., and Huttner, W. B. (1984) FEBS Lett. 177, 129-134.
4. Hortin, G., Sims, H., and Strauss, A. W. (1986) J. Biol. Chem. 261, 1786-1793.
5. Liu, M.-C., Yu, S., Sy, J., Redman, C. M., and Lipmann, F. (1985) Proc. Natl. Acad. Sci. USA 82, 7160-7164.
6. Jukkola,A., Risteli,J., Niemela, O., and Risteli, L.. (1986) Eur. J. Biochem. 154, 219-224.
7. Hortin, G., Fok, K.F., Toren, P.C., and Strauss, A.W. (1986) Submitted.
8. Lijnen, H. R., Van Hoef, B., Wiman, B., and Collen, D. (1985) Thrombosis Res. 39, 625-630.
9. Hortin, G., Tollefsen, D.M., and Strauss, A.W. (1986) J. Biol. Chem., In press.
10. Dayhoff, M. O. (1972) Atlas of Protein Sequence and Structure, Vol. 5, National Biomedical Research Foundation, Washington, D.C.
11. Brand, S.J., Klarlund, J., Schwartz, T.W., and Rehfeld, J.F. (1984) J. Biol. Chem. 259, 13246-13252.Chou, P.Y. and Fasman, G.D. (1978) Ann. Rev. Biochem. 47, 251-276.
12. Chou, P.Y. and Fasman, G.D. (1978) Ann. Rev. Biochem. 47, 251-276.
13. Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105-132.
14. Kernighan, B.W. and Ritchie, D.M. (1978) in The C Programming Language, Prentice-Hall Inc.
15. Vargas, F., Frerot, O., Dan Tung Tuong, M., and Schwartz, J.C. (1985) Biochemistry 24, 5938-5943.
16. Huttner, W., Baeuerle, P.A., Benedum, U., Friederich, E., Hille, A., Lee, R.W.H., Rosa, P., Seidel, U., and Suchanek, C. (1986) in Hormones and Cell Regulation, pp. 199-217 (Ed.: J. Nunez et al) Colloq. INSERM Vol. 139, John Libbey Eurotext, Ltd.

17. Hortin, G., Chan, A.C., Fok, K.F., Strauss, A.W., and Atkinson, J.P. (1986) J. Biol. Chem. 261, 9065-9069.
18. Belt, K.T., Carroll, M.C., and Porter, R.R. (1984) Cell 36, 907-914.
19. Ragg, H. (1986) Nucleic Acids Res. 14, 1073-1088.
20. Morita, T. and Jackson, C. (1986) J. Biol. Chem. 261, 4008-4014.
21. Hung, M.-C and Wensink, P.C. (1983) J. Mol. Biol 164, 481-492.
22. Dodt, J., Muller, H.P., Seemuller, U., and Chang, J.-Y. (1984) FEBS Lett. 165, 180-184.
23. Chung, D.W., Rixon, M.W., MacGillivray, R.T.A., and Davie, E.W. (1981) Proc. Natl. Acad. Sci. USA 78, 1466-1470.
24. Boel, E., Vuust, J., Norris, F., Norris, K., Wind, A., Rehfeld, J.F., and Marcker, K.A. (1983) Proc. Natl. Acad. Sci. USA 80, 2866-2869.
25. Bentley, P.H., Kenner, G.W., and Sheppard, R.C. (1966) Nature 209, 583-585.
26. Richter, K., Aschauer, H., and Kreil, G. (1985) Peptides 6, Suppl 3, 17-21.
27. Eng, J., Gubler, U., Raufman, J.-P., Chang, M., Hulmes, J.D., Pan, Y.-C.E., and Yalow, R.S. (1986) Proc. Natl. Acad. Sci. USA 83, 2832-2835.
28. Gubler, U., Chua, A.O., Hoffman, B.J., Collier, K.J., and Eng, J. (1984) Proc. Natl. Acad. Sci. USA 81, 4307- 4310.