ARTICULOS SOBRE TECNICAS

THE COMPOSITIONAL ACCURACY IN AMINO ACID ANALYSIS: CONSEQUENCES FOR ITS USE IN PROTEIN QUALITY CONTROL

Lila Rosa Castellanos-Serra

Division of Physical Chemistry, Center for Genetic Engineering and Biotechnology, P.O. Box 6162, 10600, La Habana, Cuba. Fax: (53-7) 21 80 70

Recibido en Agosto de 1992. Aprobado en Septiembre de 1992.

Key words: amino acid analysis, quality control.

SUMMARY

Protein amino acid analysis is generally requested for product acceptance in biotechnological industry. In a recent study it was concluded that the average accuracy of the method is about 86%. Consequences of this result for using amino acid analysis as a quality control (QC) criterion is explored here by evaluating the percentual deviation in amino acid composition introduced by an arbitrarily chosen hypothetical single contaminant for several proteins. It was found that significant amounts of contaminants in a protein sample can introduce percentual deviations in the amino acid composition lower than 15%, so that no evidences of sample contamination can be deduced from the analysis. For example, human insulin can contain up to 13% of its precursor porcine insulin, 5% proinsulin, 12% des-(B23-B30) insulin or 20% des-B30 insulin; recombinant streptokinase can contain up to 15% human plasminogen, which is used in its industrial purification by affinity chromatography. The present study emphasizes the limitations to use amino acid analysis results for arguing protein purity in QC, at the present state-of-the-art of the method.

RESUMEN

La industria biotecnológica generalmente exige la realización de análisis de aminoácidos como requisito en el control de la calidad de proteínas. Un estudio reciente demostró que la exactitud promedio en los análisis de aminoácidos es de un 86%. En el presente trabajo se exploran las implicaciones de este resultado para la utilización de la técnica como criterio en el control de la calidad, mediante la evaluación de la desviación en la composición de aminoácidos de una proteína, causada por la presencia de un contaminante hipotético escogido arbitrariamente. Se encontró que una muestra de proteína puede contener cantidades significativas de contaminantes que introducen desviaciones porcentuales en su composición de aminoácidos inferiores al 15%, de tal modo que no es posible deducir del análisis niguna evidencia de contaminación de la muestra. Por ejemplo, la insulina humana puede contener hasta 13% de su precursora insulina porcina, 5% de proinsulina, 12% de des-(B23-B30)insulina o 20% de des-B30 insulina; la estreptoquinasa recombinante puede contener hasta 15% de plasminógeno humano, que es utilizado en su purificación industrial mediante cromatografía de afiniddad. El presente estudio destaca las limitaciones del uso de los resultados de análisis de aminoácidos para argumentar la pureza de proteinas en control de calidad, considerando el estado actual de desarrollo del método.

INTRODUCTION

Amino acid analysis is one of the oldest automated tools in biochemistry as it has been available since the early 50's. It is an usual requisite for product acceptance in biotechnological industry, and it is a required control for many recombinant proteins. In a recent study carried out by the Association for Biomolecular Resource Facilities (ABRF) for the evaluation of amino acid analysis services, in which 41 North-American laboratories participated, it was concluded that the average accuracy of the method was about 86% and the average variability was about 7%. (1).

This paper studies the adequacy of amino acid analysis for evaluating protein purity, by determining the deviation in the amino acid composition introduced by a contaminant into a protein sample

EXPERIMENTAL PROCEDURES

A simple equation was deduced which describes the deviation in the amino acid composition of a protein when varying amounts of a known contaminant are present in the sample. For a two-component sample formed by a major component P1 and a contaminant P2, the deviation δ in the amino acid content for each residue, due to the presence of the contaminant and referred to its content in the major component can be described by:

$$\delta = \{ [a_1C/M_1 + a_2(1-C)/M_2] / (a_1/M_1)\} - 1; \quad (I)$$

where a_1 and a_2 are the number of amino acid residues per mole in the major component and in the contaminant, C and (1-C) are, respectively, the fraction of the sample weight corresponding to the major component and to the contaminant. This expression can be written as:

$$\delta(\%) = [C + (1-C)R - 1] * 100; \quad (II)$$

where $R = [a_2/M_2 * M_1/a_1]$ is a relative abundance factor for each residue in both proteins. This factor determines the influence of a contaminant on the amino acid composition.

RESULTS

Equation (II) was used to determine the influence of a single known contaminant on the amino acid composition of a protein sample. Values were calculated for a sample weight corresponding to one nanomole of protein containing variable amounts of a contaminant. If the experimental error of the method (E) is taken as 14% (1), then analyses giving deviations from the expected amino acid composition within this range ($\delta \leq E$) could be considered as acceptable. The following examples are presented to demonstrate the influence of a contaminant on the amino acid composition of a sample.

Example 1: Structurally unrelated proteins

In order to determine the effect of known contaminants on the amino acid analysis, α-2 interferon (α-IFN, M.W. 19200 Da) was arbitrarily chosen. As single contaminants superoxide dismutase (SOD, 15900 Da), streptokinase (SK, 47300 Da), human interleukin-2 (IL-2, 15500 Da), human γ interferon (γ-IFN, 16900 Da), hen egg lysozyme (14000 Da) , yeast ubiquitin (8500 Da) and human insulin (5700 Da) were considered. Table 1 shows the amino acid composition of 19 μg (1 nmole) of α-2 IFN and the resulting composition when the same sample contains a single contaminant. For an experimental error of 14% (1), it was found that significant amounts of some contaminants (up to 20% of Il-2, or up to 10% of γ-IFN or streptokinase) originate for all residues deviations lower than the experimental error of the method (figure 1).

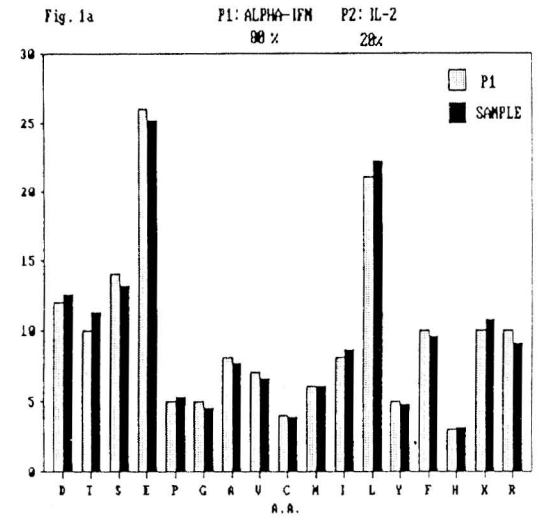
For other contaminants, only one or a small number of residues are significantly altered. For lysozyme, δ Gly = 21%, for ubiquitin, δ Gly = 17%. When the content of lysozyme and ubiquitin are 6%

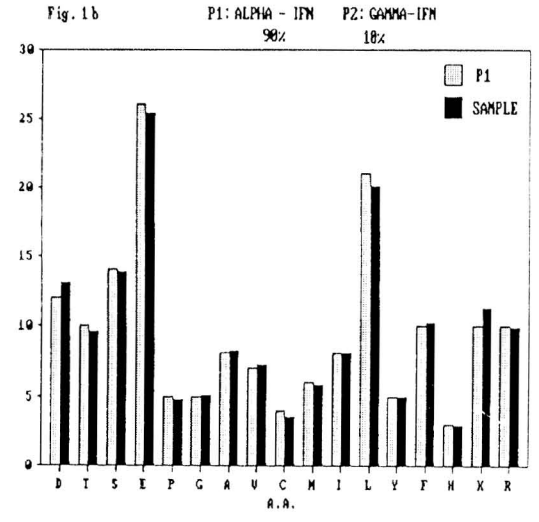
Table 1
Effect of a known contaminant on the amino acid composition of a protein sample

Amino	α2	+ 10	%	+ 10	%	+ 10	%	+ 20	%	+ 10	%	+ 10	%	+ 10	%
Acid	IFN	SOD	δ	SK	δ	γIFN	δ	IL-2	δ	Lysoz	δ	Ubiq	δ	Ins	δ
D	12	12.96	8.00	13.39	11.56	13.04	8.93	12.54	4.52	13.55	12.93	12.59	4.90	11.80	-1.67
T	10	9.96	-0.40	10.25	2.53	9.56	-4.32	11.19	11.87	9.92	-0.83	10.56	5.65	10.00	0.00
S	14	13.80	-1.43	13.57	-3.07	13.85	-1.07	13.16	-5.99	13.91	-0.64	13.72	-2.02	13.60	-2.86
E	26	24.96	-4.00	25.22	-3.00	25.44	-2.13	25.21	-3.03	24.06	-7.48	25.86	-0.54	25.73	-1.03
P	5	5.10	2.00	5.39	7.79	4.72	-5.46	5.23	4.52	4.76	-4.76	4.95	-1.06	4.83	-3.33
G	5	7.50	50.00	5.31	6.17	5.06	1.36	4.49	-10.19	6.07	21.56	5.84	16.82	5.83	16.67
A	8	8.40	5.00	8.05	0.61	8.11	1.36	7.63	-4.68	8.77	9.66	7.42	-7.21	7.53	-5.83
V	7	7.98	14.00	7.23	3.28	7.21	2.98	6.58	-5.99	7.09	1.23	7.19	2.77	7.63	9.05
C	4	4.08	2.00	3.60	-10.00	3.60	-10.00	3.94	-1.61	4.65	16.30	3.60	-10.00	5.62	40.00
M	5	4.62	-7.60	4.70	-5.94	5.07	1.36	5.24	4.77	4.76	-4.74	4.73	-5.48	4.50	-10.00
I	8	8.28	3.50	8.13	1.67	8.00	0.00	8.61	7.58	7.99	-0.17	8.76	9.56	7.87	-1.67
L	21	19.98	-4.86	20.48	-2.46	20.04	-4.59	22.19	5.68	19.95	-5.01	20.91	-0.42	20.90	-0.48
Y	5	4.50	-10.00	5.39	7.86	4.95	-0.91	4.74	-5.29	4.89	-2.14	4.72	-5.53	5.83	16.67
F	10	9.48	-5.20	9.61	-3.91	10.14	1.36	9.47	-5.29	9.39	-6.07	9.45	-5.53	10.00	0.00
H	3	3.66	22.00	3.10	3.53	2.93	-2.43	3.14	4.52	2.83	-5.63	2.92	-2.55	3.37	12.22
K	10	10.32	3.20	10.29	2.99	11.27	12.72	10.70	7.25	9.74	-2.14	10.56	5.66	9.33	-6.67
R	10	9.48	-5.20	9.73	-2.69	9.91	-0.91	8.98	-10.19	10.44	4.41	9.89	-1.06	9.33	-6.67

Lysoz: Lysozyme, Ubiq: Ubiquitin, Ins: Insulin

Note: AA: amino acid residue. α-2 IFN: amino acid composition of one nanomole of the protein. The other colums describe the amino acid composition of the same sample weight when a contaminant is present. At the column heading is the per cent of contaminant in the sample expressed as weight





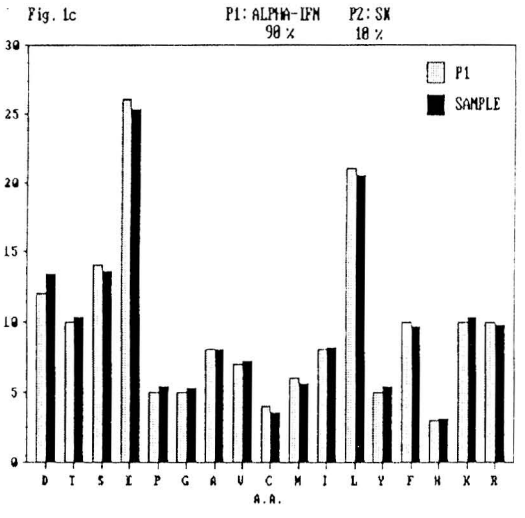


Fig .1 Bar representation of the amino acid composition of α-2 IFN and of the same sample weight when containing 20%IL-2 (1a), 10% γ-IFN (1b) or 10% SK (1c).

and 8% respectively, no evidence of sample contamination can be deduced from the analysis.

For a sample of α-2 IFN, containing 10% SOD or 10% insulin, a few residues are significantly altered: for 10% SOD, Gly (δ = 50%), Val (δ = 14%) and His (δ = 22%) are affected; for 10% insulin, Gly (δ = 17%), Cys (δ = 40%) and Tyr (δ = 17%) are affected. When the content of SOD and insulin in the sample is reduced to 5%, only one δ value is outside the experimental error range (for 5% SOD, Gly: 5/6.25, δ = 25%, for 5% insulin, Cys: 4/4.81, δ = 20%).

These examples illustrate that, in many cases, considerable amounts of contaminants do not introduce significant deviations in the amino acid composition of a sample, and that, in some cases, only one or few residues are significantly altered. In these cases sample contamination could not be suspected from the analysis.

Example 2: Structurally related proteins

Insulin and related products. Semisynthetic human insulin is obtained from porcine insulin by enzyme catalyzed transpeptidation. A sample of human insulin containing as much as 13% of porcine insulin, still gives an "acceptable" analysis, $\delta$for Thr is -4.3% and for Ala is 13%. (Thr: pure / contaminated =3 / 2.87; Ala: pure/contaminated = 1/ 1.13).

Biosynthetic human insulin, obtained by the expression of proinsulin in bacteria, can contain the precursor proinsulin, the cleaved C peptide, and the products of proteolytic cleavage in Chain B at Arg 22 (des- B23-B30 insulin) and at Lys 29

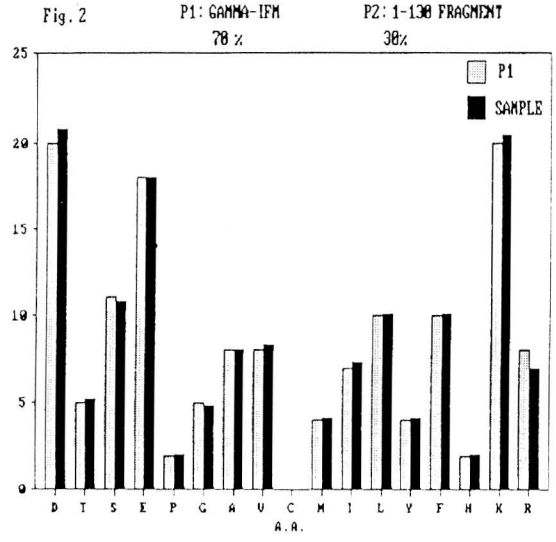


Fig.2 Bar representation of the amino acid composition of γ-IFN and of the same sample weight when containing 30% of its degradation product γ-IFN (1-130).

Table 2
Effect of contaminants on insulin amino acid analysis

Amino Acid	Insulin	+ 10 % Proinsulin	δ	+ 5 % C-Peptide	δ	+ 15 % des B23-B30	δ
D	3	2.95	-1.56	2.95	-1.83	3.38	12.58
T	3	2.89	-3.67	2.85	-5.00	2.83	-5.81
S	3	3.02	0.56	3.04	1.33	3.38	12.58
E	7	7.25	3.57	7.41	5.86	7.88	12.58
P	1	1.08	8.39	1.13	13.39	0.85	-15.00
G	4	4.30	7.42	4.47	11.63	4.23	5.69
A	1	1.15	15.33	1.24	23.50	1.13	12.58
V	4	3.98	-0.50	3.99	-0.25	4.50	12.58
C	6	5.78	-3.67	5.70	-5.00	6.75	12.58
M	0	0.00	--	0.00	--	0.00	-
I	2	1.93	-3.67	1.90	-5.00	2.25	12.58
L	6	6.16	2.67	6.27	4.50	6.75	12.58
Y	4	3.85	-3.67	3.80	-5.00	4.32	5.69
F	3	2.89	-3.67	2.85	-5.00	3.10	3.39
H	2	1.93	-3.67	1.90	-5.00	2.25	12.58
K	1	1.03	2.67	0.95	-5.00	0.85	-15.00
R	1	1.15	15.33	0.95	-5.00	1.13	12.58

des B23-B30 : des B23-B30 insulin

Note: Data are reported for a sample weight corresponding to one nanomole of human insulin

(des-B30 insulin) if extensive tryptic digestion occurs during the elimination of C-peptide. Their abundance in a sample of human insulin can be as much as 10, 5, 15% and 40% respectively, introducing deviations in the amino acid composition lower than the experimental error (table 2). For 5% C-peptide in the sample, only Ala is altered (1/ 1,23; δ = 23 %). For 40% des-B30 insulin in the sample, Thr : 3 / 2.61, δ = -13% , for all other residues, δ = 0.7%. The low efficiency of amino acid analysis for detecting insulin-related contaminants contrasts with the rigid regulatory requirements for insulin, which limit their tolerance to parts per million.

Example 3: Degradation products as contaminants

Epidermal growth factor (EGF) is usually obtained as two species, lacking respectively Arg and Arg plus Leu residues (at the C-terminus) (3,4). Natural EGF (1-53) containing up to 30% of the 1-51 specie (lacking Leu-Arg C-terminus) does not exhibit a substantial modification in the content of Leu and Arg; for Leu: 5 / 4.74 (δ = - 5.20%); for Arg: 3 / 2.72, (δ = -9.33%). Another recombinant protein, γ-IFN, is highly sensitive to proteolytic cleavage at two basic clusters (Lys-Arg rich regions) (*). It was found that γ- IFN can be accompanied by as much as 30% of its cleavage product (1-130) or 13% of its cleavage product (1-88) with no significant evidence by amino acid analysis. (figure 2).

Example 4: Process-related proteins

Table 3 presents three examples of how potential contaminants associated to protein production can influence the analysis. First, two recombinant fusion proteins are considered: superoxide dismutase-proinsulin (SOD-PI) (5) and human growth hormone fragment-β nerve growth factor (HGH-NGF) (6), both of which are precursors for proinsulin and β nerve growth factor(βNGF), respectively. The potential

(*) Pérez, L., Castellanos, L. Unpublished results

Table 3
Effect of potential process-related contaminants on the amino acid analysis of the target products

Amino		+ 7	%	+ 10	%		+ 5	%	+ 5	%		+ 15	%
Acid	Proins	SOD	δ	SOD-PI	δ	βNGF	HGH	δ	HG-NGF	δ	SK	Plasm	δ
D	4	4.45	11.31	4.42	10.46	13	12.84	-1.20	12.87	-1.02	64	60.29	-5.80
T	3	3.12	3.85	3.11	3.64	10	9.83	-1.71	9.91	-0,95	31	31.00	0.00
S	5	5.06	1.14	5.06	1.16	11	11.03	0.24	11.01	0.11	24	24.74	3.08
E	15	14.48	-3.47	14.54	-3.06	6	6.65	10.78	6.35	5.88	45	44.76	-0.54
P	3	2.99	-0,22	3.00	0.00	3	3.22	7.35	3.12	4.00	22	24.12	9.65
G	11	11.25	2.25	11.24	2.17	7	6.86	-2.06	6.92	-1.14	20	21.65	8.24
A	4	4.13	3.17	4.12	3.02	6	5.95	-0.88	5.97	-0.50	21	20.72	-1.35
V	6	6.15	2.49	6.14	2.40	13	12.56	-3.42	12.76	-1.88	23	23.11	0.49
C	6	5.74	-4.29	5.77	-3.80	6	5.74	-4.31	5.86	-2.38	0	3.72	--
M	0	0.04	--	0.04	-	2	2.02	1.17	2.01	0.63	5	5.02	0.49
I	2	2.23	11.31	2.21	10.46	6	5.99	-0.20	5.99	-0.13	23	21.18	-7.93
L	12	11.53	-3.95	11.58	-3.49	3	3.71	23.81	3.39	13.00	39	36.40	-6.66
Y	4	3.72	-7.00	3.75	-6.28	2	2.06	3.32	2.04	1.75	22	21.02	-4.44
F	3	2.95	-1.58	2.96	-1.32	7	7.02	0.29	7.01	0.14	15	14.30	-4.67
H	2	2.19	9.28	2.17	8.60	4	3.88	-2.94	3.94	-1.63	10	10.28	2.82
K	2	2.31	15.38	2.28	14.18	9	8.71	-3.17	8.84	-1.75	32	30.84	-3.62
R	4	3.88	-2.93	3.90	-2.56	7	7.02	0.29	7.01	0.14	18	18.55	3.08

Proins: Proinsulin, HG-NGF: HGH-NGF, Plasm: Plasminogen

Note: Target products: proinsulin (PI), beta-nerve growth factor (NGF) and streptokinase (SK); potential contaminants: for PI, the fusion protein SOD-PI and the cleaved fusion fragment SOD; for NGF, the fusion protein HGH-NGF and the cleaved fusion fragment HGH; for SK, human plasminogen. Data are reported for a sample weight corresponding to one nanomole of the target protein.

contaminants are the precursor fusion protein and the cleaved fusion fragment, SOD and HGH. SOD is separated by the cyanogen bromide cleavage of a methionine residue at the NH_2-terminus of proinsulin; the HGH fragment is cleaved by the action of thrombin on a tripeptide (VPR) placed at the NH_2-terminus of beta- NGF.

As a third example, recombinant streptokinase is considered: the potential contaminant is, in this case, the immobilized matrix ligand, human plasminogen (91 000 Da) used in its purification by affinity chromatography (7).

For proinsulin, 7% SOD and 10% SOD-PI in the sample introduce deviations (II) whithin the experimental range of error of the method, originating "acceptable" analyses which do not detect them; for βNGF, sample contamination can be suspected when 10% of the cleaved promoter HGH is present, as it modifies Leu and Glu significantly: Leu: 3 / 4.43, δ = 47%; Glu: 6 /7.29, δ = 21.5 %. However, when its content is reduced to 5% , only Leu is significantly altered (Leu: 3 / 3.71, δ = 23.8 %). When the fusion protein HGH-NGF is at an 8% concentration in the sample, it introduces a significant modification in only one residue (Leu, 3 /3.62, δ =21%). When HGH-NGF is present at 5 % or lower, it is no longer detectable by the analysis. Plasminogen could be present at 15% in the final product streptokinase (SK) and its presence not be suspected from the analysis, except by the detection of Cys, which is absent in SK.

DISCUSSION

Those familiar with amino acid analysis have the perception that this method is not adequate for judging protein purity. This study provides a rationale for sustaining this assessment.

Calculation of the deviation δ for some two-component protein samples has demonstrated that, in many cases, significant amounts of a contaminant introduce an insignificant deviation in the amino acid composition at the state-of-the-art of this technique. For some of the examples discussed here, an acceptable agreement with the expected values is found, except for one or two residues that are significantly altered.

Abnormally low values, resulting from protein contamination, for some specific residues in an otherwise correct analysis may not be necessarily interpreted as sample contamination, as they can be rationalized on the basis of their particular chemical behavior. An analysis showing only low recovery for serine or threonine, that are partially dehydrated under acid hydrolysis, or for an oxygen labile residue or for a

residue participating in an hydrophobic cluster, would be probably accepted as correct, if this value is attributed to its chemical behaviour. On the other hand, values beyond those accepted by the experimental error for AsX, GlX, Gly, Pro, Ala, His, Lys and Arg, which are stable under the analysis conditions and are not particularly hydrophobic, will probably warn an experienced operator of sample contamination.

Amino acid analysis is a reliable technique for determining protein concentration as well as for confirming protein identity. In these cases sample purity has to be previously demonstrated by complementary independent methods, (i.e.:chromatography, electrophoresis and immunodetection of possible contaminants).

The present paper focuses on the significance of AAA concerning QC: a "bad" amino acid analysis is meaningful as it is an evidence of sample contamination. What this study emphasizes is the meaning of a "good" analysis, as it illustrates that, with the actual accuracy of the method, a correct AAA (in which the expected and found values agree within the limits of the experimental error) is not at all an evidence of protein purity.

ACKNOWLEDGEMENTS

Thanks are due to Miriam Ribas and César Fernández for their valuable assistance in the edition of this manuscript.

REFERENCES

1.-NIECE, R. L.,L.H. ERICSSON ,A.V. FOWLER, A.J. SMITH, D.W.SPEICHER, J.W. CRABB & K.R. WILLIAMS, (1991) In: *Methods in Protein Sequence Analysis,* (Jörnvall, H., Höög, H., eds.) pp 133-141, Birkhaüser Verlag Basel

2.-CASTELLANOS, L. & R.ESTRADA, (1991) *Biotecnología Aplicada,* **8**, *174-181*

3.-NASCIMENTO, C. G., A. GYENE, S.M. HALLORAN, J. MERRYWEATHER, P. VALENZUELA, K.S. STEINER, MAISIARZ & F. R., RANDOLPH, A. (1988) *Biochemistry,* **27**: 797-802 .

4.-COUSENS L.,J.R. SHUSTER, C. GALLEGOS, KU, LAILING, M. STEMPIEN, M. URDEA, R. SANCHEZ-PESCADOR, A. TAYLOR & P. OLSON, (1987) *Gene*, **61**, 265-275

5.-OHTSUKA, E.(1989) *European Patent Application* 0 329 175 A1 (17.02.1989)

6.-RODRIGUEZ, P., L. HERNANDEZ, E. MUÑOZ, A. CASTRO, J. DE LA FUENTE,& L. HERRERA, (1992) *Bio/Techniques*, **12**, 424-429