

# 10

## A SYNTHETIC PEPTIDE FOR EVALUATING PROTEIN SEQUENCER AND AMINO ACID ANALYZER PERFORMANCE IN CORE FACILITIES: DESIGN AND RESULTS

Ronald L. Niece<sup>1</sup>, Kenneth R. Williams<sup>2</sup>,  
Cynthia L. Wadsworth<sup>1</sup>, James Elliott<sup>2</sup>,  
Kathryn L. Stone<sup>2</sup>, Walter J. McMurray<sup>3</sup>,  
Audree Fowler<sup>4</sup>, Donna Atherton<sup>5</sup>,  
Rusty Kutny<sup>6</sup>, Alan J. Smith<sup>7</sup>

<sup>1</sup> University of Wisconsin Biotechnology Center  
Protein/DNA Sequence/Synthesis Facility  
Madison, WI 53705

<sup>2</sup> Yale University School of Medicine  
Protein and Nucleic Acid Chemistry Facility  
New Haven, CT 06510

<sup>3</sup> Yale University Comprehensive Cancer Center  
New Haven, CT 06510

<sup>4</sup> Dept. of Biological Chemistry  
and Molecular Biology Institute  
UCLA School of Medicine  
Los Angeles, CA 90024

<sup>5</sup> The Rockefeller University  
New York, NY 10021

<sup>6</sup> Eastman Pharmaceuticals  
Great Valley, PA 19355

<sup>7</sup> Dept. of Biological Chemistry  
University of California  
Davis, CA 95616

## I. INTRODUCTION

Recent advances in the biotechnology of protein and nucleic acid sequencing and synthesis have fostered the development of more than 100 core facilities in the U.S. whose primary function is to bring these techniques within reach of a wide spectrum of investigators that includes not only specialists in these areas but also cellular and molecular biologists in general. Because much of the needed equipment is expensive and requires considerable expertise as well as a large number of samples to keep it operating continuously at peak efficiency, core facilities staffed by professional personnel represent an effective and economical means of bringing these technologies to bear on problems related to biochemical research. In order to efficiently utilize a core facility, a prospective user must have realistic expectations concerning the technical capabilities of these facilities. A recent survey of forty core facilities clearly demonstrates that the relatively isolated instances where extremely high sensitivity analyses have been reported are not representative of what can be expected when a typical "unknown" protein sample, prepared by the investigator, is submitted to a well-run core facility, whether it is located in a university, government, or industrial laboratory (1). As part of our plans to provide the sequencing community with control materials, the Research Resource Facilities Group designed and synthesized a 40-residue peptide with characteristics suited for evaluation of protein/peptide sequencers and amino acid analyzers. The peptide was initially released to 103 core facilities as an "unknown" sample to provide data concerning the protein sequencing and amino acid analysis capabilities of existing core facilities. Prior to revealing the actual sequence and composition of these standard peptides, STD-1 and STD-2, at the Second Symposium of the Protein Society, sixty sets of data had been returned which together provide the basis for this report. <sup>1</sup>

---

<sup>1</sup> Samples of STD-1 and STD-2 are available from KRW.

## II. MATERIALS AND METHODS

### A. Design of the STD-1 and STD-2 Peptides

The forty-residue STD-1 and STD-2 peptides designed at the University of Wisconsin Biotechnology Center (CLW and RLN) were sufficiently long to challenge the capabilities of automated protein sequencing instruments and with suitable compositions to challenge amino acid analyzer capabilities. The standard peptides were identical except that the one designed for amino acid analysis (STD-2) had the amino-terminus acetylated. The amino acid sequence of these two peptides is shown in Fig. 2.

For protein sequencing the amino-terminal portion of the peptide (STD-1) was designed to provide data reflecting sequencer performance. Proline (which frequently results in incomplete cleavage and excessive lag) and serine (which may cause some blockage of the amino-terminus) were avoided. The sequencer performance could be assessed quickly on an overnight run for repetitive yield using alanine residues at positions 4 and 10 and for initial yield using valine, tyrosine, and alanine at positions 1, 2, and 4. Alanine residues were regularly spaced throughout the sequence to permit an accurate calculation of repetitive yields over various sequences within the peptide. Many less well recovered amino acids such as aspartate, glutamate, arginine, tryptophan, histidine, and cysteine were present early in the sequence to facilitate their identification and accurate quantitation. The middle portion of the sequence was designed to provide sequence interpretation challenges due to increasing lag before washout became significant. Proline was inserted at positions 18 and 25 for this purpose. The carboxy-terminal third of the peptide was designed to present sequencing challenges in the face of increasing washout. Arginine residues at positions 38 and 40 were present to limit washout. The two serines at residues 31 and 32 followed by glycine were expected to be difficult to interpret. With the limited amount of material present, it was expected that only a few picomoles (pmol) of material would be sequencing after 30 cycles of degradation.

In addition to the features relating to protein sequencer performance, consideration was also given to HPLC identification and quantitation of the resulting PTH-derivatives. Aspartic acid was placed early in the sequence to evaluate

its separation from the ammonia-PITC artifact which is prominent in early cycles. Tryptophan was also placed early to permit observation of its oxidation products and its separation from DPU (diphenylurea). Histidine was positioned in the sequence between alanine and cysteine because PTH-histidine is usually eluted between PTH-alanine and PTH-dehydroalanine (formed from unmodified cysteine during sequencing). Similarly, arginine was positioned in the sequence between alanine and tyrosine. Histidine and arginine appeared throughout the sequence to permit monitoring the constancy of their elution times throughout the course of the run.

In terms of amino acid composition, the standard peptide was designed so that after acid hydrolysis all 18 of the commonly observed amino acids would be present. To monitor the extent of hydrolysis, a sequence consisting of isoleucyl-isoleucyl-valine was present. The ratios of isoleucine to leucine and of serine to threonine were both 3 to 1 to monitor the degree of resolution of these pairs of amino acids that are frequently incompletely resolved by ion exchange chromatography.

General sequence design considerations included ease of peptide synthesis, stability, and solubility. The peptide was short enough and of a suitable sequence to synthesize chemically and to permit rigorous characterization. Bonds potentially unstable at the extremes of pH found in the sequencer, specifically ASP-PRO and ASN-GLY, were also avoided. Tryptophan is unstable and was placed late in the peptide synthesis to minimize its destruction. A sufficient number of amino acids charged at different pH ranges were included to permit solubility in a broad range of HPLC solvent systems.

Finally the peptide was designed to be useful in monitoring a variety of other techniques typically used in a core facility. It was large enough so that it could be run on SDS polyacrylamide gels and electroblotted. Chemical and enzymatic cleavage sites were incorporated so that the resulting fragments could be sorted according to different characteristics. For example, the extent of cyanogen bromide (CNBr) cleavage could be monitored at the two methionines neither of which was followed by serine or threonine (which can reduce the yield of CNBr cleavage). One CNBr cleavage site was closely followed by proline so that the mixture of peptides produced during in situ CNBr cleavage could be sorted out by blocking non-proline terminated peptides at the appropriate sequencing cycle. Several different enzymatic cleavage sites were incorporated. The potential fragments differ in hydrophobicity and/or charges to permit HPLC separation.

## B. Synthesis, Characterization and Distribution of STD-1 and STD-2 Peptides

The STD-1 peptide was synthesized at Yale University (JE) on an Applied Biosystems Model 430A Solid Phase Peptide Synthesizer using standard techniques. After synthesis and removal of the NH<sub>2</sub>-terminal tBOC, one half of the resin was acetylated with acetic anhydride to make STD-2. Following HF cleavage, the peptide/resin mixture was extracted with 40 ml 7M guanidine hydrochloride and the peptides purified on two Vydac C-18 columns connected in series. The sequence of

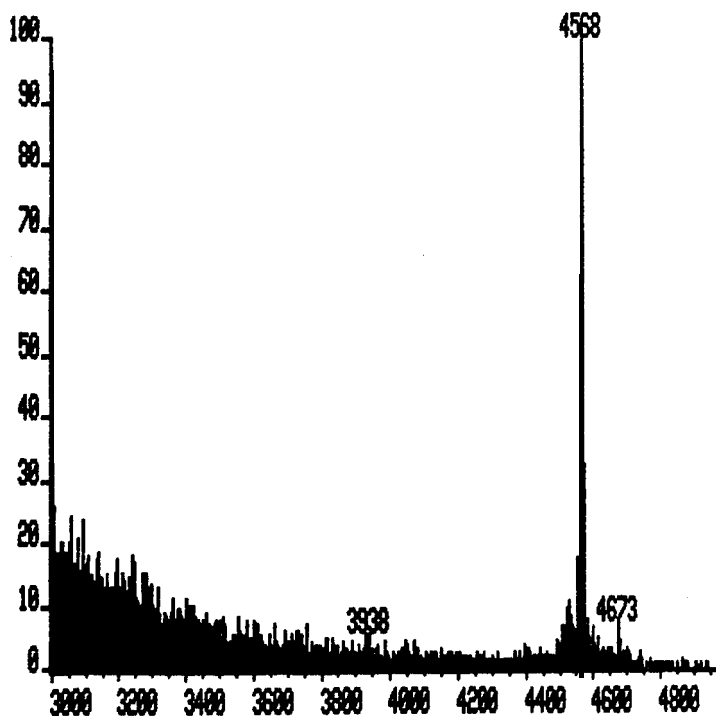


Fig. 1 FAB positive ion mass spectrometry of STD-2. Based on this analysis, the protonated molecular ion of STD-2 has a mass of 4568.22 daltons (predicted value is 4568.31 daltons). Sample was recorded using a thioglycerol matrix and a xenon gun operated at 8 keV.

STD-1 was confirmed by direct sequencing of 5 nanomoles (nmol); its purity was at least 98% with 0.9% preview. The molecular weights were determined by FAB mass spectrometry (WM). As shown in Fig. 1, the observed and predicted protonated molecular weights for STD-2 were within 0.1 atomic mass units of each other.

Aliquots of the standard peptides (100 pmol **STD-1** and 1.09 nmol STD-2) were dried in 1.5 ml Eppendorf tubes that had been pre-washed in 0.05% trifluoroacetic acid, 50% acetonitrile and then distributed to the 103 core facilities on a mailing list that had been compiled during the previous two years by the Research Resource Facilities Group. Specific instructions for re-solubilizing both the STD-1 and STD-2 peptides were included with these samples. Extensive testing of randomly selected aliquots indicated that by following the instructions it would be possible to load approximately 75 pmol of STD-1 onto the sequencer and to hydrolyze either 0.5  $\mu$ g or 4.5  $\mu$ g STD-2. The instructions specified that for high sensitivity analyzers 10% (*i.e.* 0.5  $\mu$ g) and that for less sensitive analyzers 90% (*i.e.* 4.5  $\mu$ g) of STD-2 should be taken for hydrolysis and amino acid analysis, respectively. To guarantee the confidentiality of the resulting data, the responses were returned to a third party who removed any postmarks or other identification relating to the originating laboratory prior to forwarding the data to the authors.

### III. RESULTS AND DISCUSSION

#### A. Sequencing of STD-1

A total of 54 responses were received among which two instrument failures were reported. This indicates that there is about a 4% chance of instrument failure when unknown samples are run. Fig. 2 provides a graphical representation of the 54 sets of STD-1 sequencing data. The sequence of the peptide is given at the bottom. The figure indicates the overall number of correct, incorrect, and tentative calls at each position in the sequence. The same data are grouped and averaged in Table I according to instrument model. Based on the data in Table I, Applied Biosystems instruments clearly give superior results with this peptide compared to those obtained with the Beckman 890 "spinning cup" instrument or a manual approach. Because the peptide was designed to be used with high sensitivity instrumentation, only a limited amount was provided accounting for the relatively poor showing of

## STANDARD 1 PEPTIDE

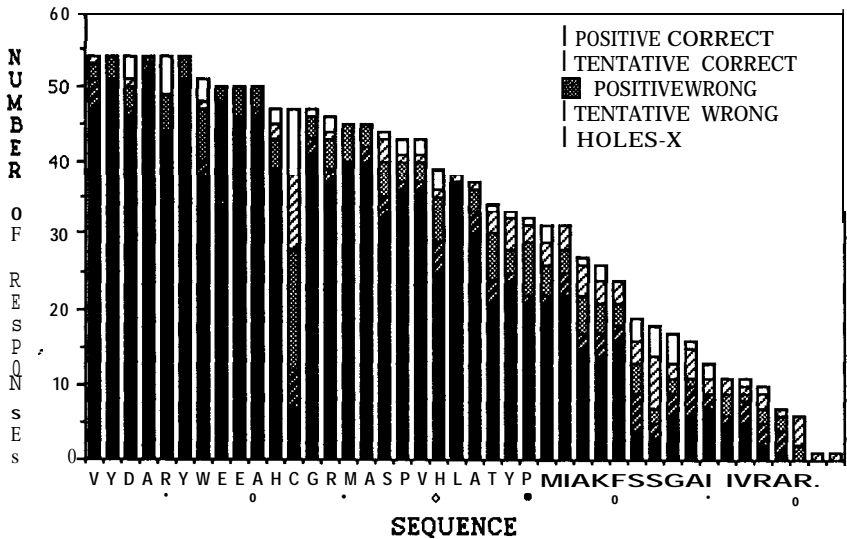


Fig. 2 Distribution of calls for each cycle. Any indication of qualification or uncertainty by the operator led to placing that amino acid in the tentative identification category. When more than one residue was listed in a cycle it was plotted as wrong. Cycles for which no residue was called and cycles labelled X were plotted as holes. The sequence is listed across the bottom with the amino terminus at the left.

older instrumentation and methods. Similarly, the sequence was selected to evaluate performance of modern instrumentation. It is also clear that the addition of the Model 120 on-line PTH-HPLC to the Model 470 sequencer approximately doubled the number of residues that can be sequenced. In contrast, there does not seem to be any significant difference in either the number of residues sequenced, the repetitive yields, or the accuracy of the three models of HPLC equipped instruments. For these three instruments the overall average number of correctly identified residues was 23.9 and the accuracy of sequence calls was 94.6% for those cycles that were "positively identified". The accuracy for the 93 residues which were only tentatively identified was 48%; this indicates a very good ability on the part of the personnel who interpreted the data to discriminate between a confident and tentative call.

TABLE I. Sequencing of STD-1

Instrument	Number	Coupling yield <sup>a</sup> (picomole)			Repetitive yield <sup>b</sup> (%)			Number of correct <sup>c</sup> Residues			Accuracy <sup>c</sup> (%)		
		Mean	Range	Std. Dev.	Mean	Range	std. dev.	Mean	Range	Std. dev.	Mean	Range	std. dev.
<b>Applied Biosystems</b>													
470A	4	44.0	--	--	90.6	87.7-93.9	2.5	12.2	6-17	4.2	92.4	75.0-100	10.3
<b>Applied Biosystems</b>													
477A/120	7	77.8	--	--	89.0	81.7-92.2	3.6	25.6	14-35	7.0	91.6	64.5-100	11.5
475A/120	5	50.6	37.6-64.3	10.9	86.7	76.3-93.0	5.6	20.8	9-32	9.5	94.9	90.0-100	4.3
470A/120	26	50.9	<u>24.1-81.7</u>	<u>15.4</u>	87.1	<u>79.6-92.7</u>	3.4	24.0	<u>10-37</u>	6.4	95.4	<u>79.3-100</u>	5.5
Total	38	52.5	24.1-81.7	15.6	87.5	76.3-92.7	3.9	23.8	9-37	7.1	94.6	64.5-100	7.0
<b>Beckman</b>													
890	3	--	--	--	--	--	--	2.0	0-3	1.4	30.9	0-75.0	32.0
<b>Manual</b>													
	2	--	--	--	--	--	--	2.5	0-5	2.5	41.7	0-83.3	41.7

<sup>a</sup> Coupling yield from the data sets in which the recovery of valine in cycle 1 could be calculated unambiguously. The total number of such cases was 477/120 (1), 475/120 (3), 470/120 (12) and 470 (1).

<sup>b</sup> Based on the background corrected yield of alanine 4 and alanine 16 or, if the sequence did not extend this far, on the yield of alanine 4 and alanine 10.

<sup>c</sup> Based on positively identified cycles only; those cycles that were left blank or that had either multiple or tentative assignments were not scored.

In addition to the sequencing data above, results were also obtained from three laboratories that specialize in extended sequencing runs on low picomole amounts of samples. As shown in Table II, these laboratories correctly identified an average of 36.5 residues with an accuracy of 99%. The average repetitive yield was greater than 91%. The instruments used were Models 470A and 477A equipped with the Model 120 on-line PTH-HPLC.

Five of the twelve respondents with Model 475A and 477A sequencers included the data generated from the computer assisted automatic sequence identification. Although the version of software was not specified, the runs were initiated at least 2 months after the release of version 1.30 for 900A and 1.50 for 477A in January 1988. Sequence calling routines were identical to those in the previous versions (1.10 and 1.20 for 900A and 477A, respectively) which were released in May and June 1987. Comparison of the number of errors in the sequence identified by the operator with those in the computer generated sequence to the end of confident operator calls, a total of 97, indicates that unreviewed use of the

TABLE II. Results of STD-1 sequencing on three "high performance" instruments

	470A/120	477A/120	477A/120
Positively identified residues			
correct	38	34	37
incorrect	0	1	0
% accuracy	100	97	100
Tentatively identified residues			
correct	--	3	--
incorrect	--	1	--
Data system calls through last manual call			
correct	--	34	30
incorrect	--	5	9
holes	--	1	0
% accuracy	--	87	77
Data system calls through last picomole ratio greater than 10			
correct	--	30	29
incorrect	--	2	1
holes	--	1	0
% accuracy	--	94	97
Repetitive yield	92	91	91
First cycle yield	47	57	77

automatic sequence identification would produce sequences of consistently lower accuracy, only 89% compared to 95%. More errors occurred when the picomole ratio (a software defined statistic relating amount of the identified amino acid to its background amount) was small. If automatic calls only to the last cycle with the picomole ratio greater than 20 (rather than through the last operator call) were considered, the accuracy was the same as confident operator calls (95%), but only on 56 compared to 97 calls. If data through the last cycle with picomole ratio greater than 10 were considered, the accuracy was 90% of 72 calls and can be defined as the end of the confident computer calls. Within the data set of confident computer calls for picomole ratios between 11-20 there were 12 correct out of 16 calls; for picomole ratios between 5-10 there were 10 correct out of 12 calls; and for picomole ratios less than 5 there were 2 correct out of 5 calls. Results from labs specializing in extended sequencing were similar and are included in Table II. Clearly operators identify sequences farther and more accurately than the computer; however, this analysis suggests that when the operator and the computer calls differ within the limit of confident computer calls the cycle merits re-evaluation by the operator.

#### B. Amino acid analysis of STD-2

A total of 48 responses included amino acid composition data, and of these, three (6%) reported instrument malfunctions. Table III summarizes the other 45 compositions. The data were standardized by calculating the mole percent of each amino acid multiplied by 40 (the length of the peptide) to obtain the residues per mole of peptide. The data were divided into two groups based on the two most commonly used chemistries, post-column ninhydrin and pre-column phenylisothiocyanate derivatization. Columns 4-6 of Table III detail the results from samples analyzed using ninhydrin after hydrolysis of 90% (900 pmol) of STD-2. When 10% (100 pmol) of the sample was used for the ninhydrin method, the results were too variable to include here. The PTC analysis data, shown in columns 7-9, resulted from the hydrolysis of only 10% of the sample. Too few results of fluorescent detection methods (OPA, Fmoc) were received to permit analysis.

With these data, we can determine the accuracy of an amino acid analysis of an unknown sample submitted to a core facility. In 27 of the 48 compositions (58%) submitted, the error was greater than 10%, i.e., more than 4 residues per

mole were incorrect (cysteine and tryptophan were excluded given the recognized difficulties of hydrolysis). Only eight of the data sets were better than 95% correct (2 or fewer incorrect residues). It is obvious that significant problems exist with amino acid analysis.

Interpretation of the amino acid analysis data and assessment of the possible contributing factors to these poor results was complicated by the many variables involved. Factors affecting amino acid analysis include the possible introduction of contaminants, hydrolytic destruction or incomplete hydrolysis, type of chemistry used, and instrument performance. We had attempted to eliminate hydrolysis as a variable; however, a significant fraction of analysts chose to use their own hydrolysis procedure rather than the requested protocol. After grouping the samples among all the

Table III. Amino acid analysis of STD-2

Amino acid	All responses (n=45)		Ninhydrin responses (n=10)			PTC responses (n=19)			Actual
	Mean	Std dev	Mean	Std dev	Range <sup>a</sup>	Mean	Std dev	Range <sup>a</sup>	
ALA	6.98	1.24	6.7	1.14	5.2-8.0	7.0	0.99	5.4-8.1	7
ARG	4.01	0.95	3.6	0.34	3.6-4.4	4.3	0.87	2.6-4.9	4
ASX	1.44	0.55	1.5	0.18	1.1-1.9	1.3	0.26	1.0-1.5	1
CYS <sup>b</sup>	0.54	0.29	Not calculated						1
GLX	2.65	0.81	2.7	0.18	2.3-3.3	2.5	0.65	1.8-3.6	2
GLY	3.04	0.93	3.2	0.50	2.6-4.2	2.7	0.37	2.1-3.7	2
HIS	1.89	0.45	1.8	0.05	1.7-2.1	1.8	0.12	1.3-2.1	2
ILE	2.27	0.49	2.5	0.26	2.0-2.9	2.2	0.24	1.4-3.0	3
LEU	1.37	0.47	1.3	0.09	1.1-1.5	1.4	0.34	1.1-1.8	1
LYS	1.26	0.31	1.2	0.10	0.9-1.6	1.2	0.05	1.0-1.4	1
MET	1.49	0.48	1.3	0.46	0.5-2.0	1.5	0.18	0.8-1.8	2
PHE	1.11	0.17	1.0	0.01	0.9-1.1	1.1	0.05	1.0-1.4	1
PRO	1.98	0.51	2.4	0.29	1.7-2.3	2.1	0.07	1.9-2.5	2
SER	3.24	0.44	2.9	0.17	2.5-3.3	3.3	0.18	2.8-3.8	3
THR	1.28	0.45	1.4	0.10	1.0-1.7	1.3	0.08	1.0-1.5	1
TYR	2.60	0.69	2.7	0.07	2.5-3.0	2.7	0.33	2.1-3.1	3
VAL	2.88	0.44	2.8	0.17	2.3-3.0	3.0	0.09	2.7-3.5	3
TRP <sup>c</sup>	1.10	0.64	Not calculated						1

<sup>a</sup> Highest and lowest values were excluded from range.

<sup>b</sup> n=11

<sup>c</sup> n=4

various categories, the sample size per group was too small to permit adequate comparisons. However, an examination of the eight "acceptable" compositions with fewer than two incorrect residues revealed no correlation with instrument type, chemistry, amount of sample, or hydrolysis conditions. Contaminants introduced during handling of the peptide could explain the poor compositional data obtained. Some facilities appear not to have sufficient controls to detect suboptimal instrument performance or the presence of contaminants; nine of the 48 compositions incorrectly quantitated 12 or more residues. Further work is needed to establish good standards and controls encompassing all phases of the amino acid analysis process used in core facilities.

#### IV. CONCLUSIONS

Based on the results obtained on STD-1, it appears reasonable to expect an average facility equipped with an Applied Biosystems protein/peptide sequencer with an on-line HPLC to correctly sequence approximately 24 residues of an unknown peptide if the peptide is at least 95% pure and at least 75 pmol is loaded onto the instrument and it is free of other primary amines and interfering substances. Under these conditions the overall accuracy of operator sequence identification is about 95% which is above the accuracy of automatic sequence identification that 6 facilities obtained with the current version of software available on the Model 900 Data System. Automatic calls extending as far as operator calls were only 89% accurate while automatic calls only to where the picomole ratio fell below 10 were 90% accurate. Although three laboratories that specialize in obtaining extended sequences from picomole amounts of peptides and proteins were able to accurately call an average of 12 additional residues of sequence, this potential cannot in general be routinely reached in a typical, high throughput core facility.

It is clear that amino acid analysis at moderate to high sensitivities lacks precision and consistency. Few laboratories can routinely provide compositions which are at least 90% correct with 0.1 to 1 nanomole of peptide.

## V. ACKNOWLEDGEMENTS

The cooperation of all of the anonymous laboratories that graciously used their resources to analyze the samples and provide the data described here was essential to making this project possible. Their support is most gratefully acknowledged. The three high performance laboratories that were not a part the regular anonymous sample are thanked for contributing their data. The assistance of Dr. Bernard Forget, Associate Dean, Yale University in providing the anonymity of the respondents is appreciated. Directors of protein and/or nucleic acid chemistry facilities desiring to be included on this mailing list of core facilities should provide their names to KRW.

## VI. REFERENCE

1. Williams, K.R., Niece, R.L., Atherton, D. Fowler, A.V., Kutny, R., Smith, A.J. (In press) FASEB J.