Direct MALDI analysis of naturally cleaved human saliva samples: Mapping to a series of KPQ-terminated peptides from small salivary proteins.

Kenneth C. Parker[1]; Na Tian[2]; Frank Oppenheim[2]; Eva Helmerhorst[2] . [1]*SimulTof Systems, Sudbury , MA;* [2]*Boston University School of Dental Medicine, Boston, MA*

TP18 #270

Introduction

One of the most easily collected human biofluids is saliva. The dominant intact proteins in saliva are usually alpha-amylase, immunoglobulin A, and lysozyme, but saliva also commonly contains naturally processed peptides in the 600- 10000 m/z range that can directly monitored by MALDI-MS. Previous experiments have established that many of these peptides derive from seven additional proteins that are highly expressed in saliva: basic salivary proline rich proteins 1-4 (PRB1-PRB4), salivary acidic proline-rich phosphoprotein (PRPC) and histatin-3 (His3). Some of the responsible proteases apparently derive from commensal bacteria, for example, Rothia species, that often cleave proteins C-terminal to the tripeptide sequence KPQ (Helmerhorst et al., 2008).

Methods

1. Collect whole or parotid secretion saliva from 88+ human subjects (BU) or lab personnel (Sudbury).
2. Spin; keep supernatant.
3. Dilute into HCCA MALDI matrix; spot in duplicate.
4. Collect MALDI reflectron MS spectra (14.8 m flight tube).
4. Map to:
   - list of 338 identified peptides
   - to series of staggered peptides (staggered PMF)

   from13 small salivary proteins.
5. Prepare 1 amu mass matrix from top 40 masses from 179 spectra from 88 patients found >=4 times -> 252 masses.
6. Normalize, perform PCA.
7. **Collect selected MSMS spectra.**

Staggered PMF

1. Get protein sequence of salivary protein
2. Make truncated peptide series starting at every possible N-terminus and at every possible C-terminus (each peptide ends up in 2 series).
3. Define each series of related peptides as a protein-like entity for PMF.
4. Increase ChemScore of peptides 2x for C-ter. Q and N-ter. G.
5. Use ordinary PMF logic to identify those series that are most prominent (based on Parker(2002)).

Table. 1. Example truncation series from histatin 3 (His3)

Shared mature N-terminus Shared C-terminus at aa 24

aa	Mass	<	Sequence	>	mb
1	987.5		DSHAKRHH	GYK	987
1	1044.5		DSHAKRHHG	YKR	1044
1	1207.6		DSHAKRHHGY	KRF	1207
1	1335.7		DSHAKRHHGYK	RKF	1335
1	1491.8		DSHAKRHHGYKR	KFH	1491
1	1619.9		DSHAKRHHGYKRK	FHE	1619
1	1766.9		DSHAKRHHGYKRKF	HEK	1766
1	1904.0		DSHAKRHHGYKRKFH	EKH	1903
1	2033.0		DSHAKRHHGYKRKFHE	KHH	2032
1	2161.1		DSHAKRHHGYKRKFHEK	HHS	2160
1	2298.2		DSHAKRHHGYKRKFHEKH	HSH	2297
1	2435.2		DSHAKRHHGYKRKFHEKHH	SHR	2434
1	2522.3		DSHAKRHHGYKRKFHEKHHS	HRQ	2521
1	2659.3		DSHAKRHHGYKRKFHEKHHSH	RQY	2658
1	2815.4		DSHAKRHHGYKRKFHEKHHSHR	QYR	2814
1	2872.5		DSHAKRHHGYKRKFHEKHHSHRG	YRS	2871
1	3035.5_		DSHAKRHHGYKRKFHEKHHSHRGY	RSN	3034

Pale blue series Fig. 3

aa	Mass	<	Sequence	>	mb
1	3035.5		DSHAKRHHGYKRKFHEKHHSHRGY	RSN	3034
2	2920.5	D	SHAKRHHGYKRKFHEKHHSHRGY	RSN	2919
3	2833.5	DS	HAKRHHGYKRKFHEKHHSHRGY	RSN	2832
4	2696.4	DSH	AKRHHGYKRKFHEKHHSHRGY	RSN	2695
5	2625.4	SHA	KRHHGYKRKFHEKHHSHRGY	RSN	2624
6	2497.3	HAK	RHHGYKRKFHEKHHSHRGY	RSN	2496
7	2341.2	AKR	HHGYKRKFHEKHHSHRGY	RSN	2340
8	2244.1	KRH	HGYKRKFHEKHHSHRGY	RSN	2243
9	2067.1	RHH	GYKRKFHEKHHSHRGY	RSN	2066
10	2010.0	HHG	YKRKFHEKHHSHRGY	RSN	2009
11	1847.0	HGY	KRKFHEKHHSHRGY	RSN	1846
12	1718.9	GYK	RKFHEKHHSHRGY	RSN	1718
13	1562.8	YKR	KFHEKHHSHRGY	RSN	1562
14	1434.7	KRK	FHEKHHSHRGY	RSN	1434
15	1287.6	RKF	HEKHHSHRGY	RSN	1287
16	1150.6	KFH	EKHHSHRGY	RSN	1150
17	1021.5	FHE	KHHSHRGY	RSN	1021

Pink series Fig. 3

aa	Mass	<	Sequence	>
75	1731.8684	QGK	PQGPPQQGGPPPPPQGR	PQG
76	1731.8684	GKP	QGPPQQGGPPPPPQGRP	QGP
77	1731.8684	KPQ	GPPQQGGPPPPPQGRPQ	GPP
78	1731.8684	PQG	PPQQGGPPPPPQGRPQG	PPQ
79	1731.8684	QGP	PQQGGPPPPPQGRPQGP	PQQ
80	1731.8684	GPP	QQGGPPPPPQGRPQGPP	QQG
81	1731.8684	PPQ	QGGPPPPPQGRPQGPPQ	QGG
82	1731.8684	PQG	GGPPPPPQGRPQGPPQG	GGH
83	1731.8684	QGG	GPPPPPQGRPQGPPQGG	GHP
84	1731.8684	GGG	PPPPPQGRPQGPPQGGG	HPR
85	1731.8684	GGH	PPPPQGRPQGPPQGGGH	PRP
86	1731.8684	GHP	PPPQGRPQGPPQGGGHP	RPP

Complication of truncation series informatics: repeat sequences

Fig. 1. Software engineer's Saliva.

Rank	MassAvg	ppm	<	Sequence	>	Chl
19	2625.4	5.4	SHA	KRHHGYKRKFHEKHHSHRGY	RSN	20
17	2341.2	-4.5	AKR	HHGYKRKFHEKHHSHRGY	RSN	20
13	1847.0	0.0	HGY	KRKFHEKHHSHRGY	RSN	20
11	1718.9	1.5	GYK	RKFHEKHHSHRGY	RSN	20
9	1562.8	0.3	YKR	KFHEKHHSHRGY	RSN	20
10	1434.7	1.2	KRK	FHEKHHSHRGY	RSN	20
3	1287.6	1.3	RKF	HEKHHSHRGY	RSN	20
5	1150.6	0.9	KFH	EKHHSHRGY	RSN	20
25	987.5	-1.6	_	DSHAKRHH	GYK	40
4	1207.6	-1.2	_	DSHAKRHHGY	KRK	20
2	1335.7	-0.9	_	DSHAKRHHGYK	RKF	20
7	1491.8	1.6	_	DSHAKRHHGYKR	KFH	20
107	1619.8	-9.2	_	DSHAKRHHGYKRK	FHE	20
76	1766.9	6.3	_	DSHAKRHHGYKRKF	HEK	20
115	2522.3	-4.9	_	DSHAKRHHGYKRKFHEKHHS	HRQ	20
51	990.5	-3.1	KPQ	GPPPQGGRPQ	GPP	100
71	1866.9	-1.3	KPQ	GPPPQGGRPQGPQGQPQ	_	160
42	1403.7	-9.0	KSR	SARSPPGKPQGPPQ	QEG	40
1	2185.1	-6.1	KSR	SARSPPGKPQGPQEGKNKPQ	GPP	80
94	1731.9	1.8	KPQ	GPPQQGGHPPPPQGRPQ	GPP	120
28	2490.2	5.4	KPQ	GPPQQGGHPPPPQGRPQGPQQGGH	PHP	80
26	1007.5	1.3	RKF	HEKHHSHR	GYK	40
3	1287.6	1.3	RKF	HEKHHSHRGY	RSN	20
15	1443.7	0.0	RKF	HEKHHSHRGYK	SNY	20
14	925.5	-3.1	KAR	HHGYKRK	FHE	20
125	1605.8	8.6	KAR	HHGYKRKFHEKH	HSH	20
108	1965.0	0.2	KAR	HHGYKRKFHEKHHSH	RGY	20
40	2121.1	0.7	KAR	HHGYKRKFHEKHHSHR	GYR	40
17	2341.2	-4.5	AKR	HHGYKRKFHEKHHSHRGY	RSN	20
128	2065.0	3.9	NKS	qSARSPPGKPQGPPQGGNQP	QG	20
43	2193.1	4.6	NKS	qSARSPPGKPQGPPQGGNQPQ	G	80
123	1864.0	7.2	PPP	qEGNKSRSARSPPGKPQG	PPQ	20
24	2186.1	-5.4	PPP	qEGNKSRSARSPPGKPQGPPQ	QEG	40
11	1718.9	-1.5	YKR	EYHEKHHSHRGY	SNY	20
15	1443.7	0.0	RKF	HEKHHSHRGYK	SNY	20
49	1306.7	0.1	KFH	YKHSHRGYR	SNY	20
64	1040.5	-4.3	HEK	HHSHRGYR	SNY	20
78	1132.5	1.1	_	DSHEKHHSHRG	YRR	20
33	1421.7	-7.1	_	DSHEKHHSHRGYR	RKF	20
70	972.6	4.8	HGY	KRKFHEK	HHS	20
81	1109.6	-6.8	HGY	KRKFHEKH	HSH	20
21	1847.0	0.5	HGY	FRKFHEKHHSHRGY	RSN	20

I	Symb	Series	Leng	#Pep	#Obs	#Obs_i	Same	Score	%IM	ppw
				51	80	18	18	12 1030002	32.8	1.3
1	His3			51	80	18	18	12 1030002	32.8	1.3
2	PRPC			166	104	6	6	4 9687	10.4	0.3
	all					24		16		
1	His3	C13			16	8	8	5 661766	15.7	1.4
2	His3	N1			25	8	8	5 412745	16.6	1.4
3	His3	N15			12	2	3	2 16369	1.3	0.5
4	His3	N7			20	4	5	0 10867	1.1	2.5
5	PRPC	N132			10	2	2	2 5024	0.4	2.4
6	PRPC	N107			36	2	2	2 4702	9.9	0.2
	all					26		16		

Mapping of previously identified peptides to Fig. 1, and Staggered series statistics. The "Same" columns lists how many members of each StaggeredPMF series were previously identified.

Table. 2. Statistics of finding mass amongst the 88 x 2 spectra

#spec	mb	Intensity	Symb	Sequence	MassAvg	SDev	ppm
80	990	73889	PRPC	GPPPQGGRPQ	990.5	0.003	-5.3
57	1004	24140			1004.5	0.016	
51	1076	21377			1076.5	0.006	
61	1135	33208			1135.6	0.010	
52	1193	21275			1193.6	0.017	
83	1224	89317	PRPC	GPPQQGGHRPP	1224.6	0.007	-2.3
116	1287	55418	His3	HEKHHSHRGY	1287.6	0.006	-1.9
59	1315	41393	SMK38	GPQRGPPPPPAPP	1315.7	0.008	-1.4
118	1335	62536	His3	DSHAKRHHGYK	1335.7	0.011	-0.3
51	1380	24307			1380.7	0.007	
85	1390	62839	PRPC	GGRPQGPPQGQSPQ	1390.7	0.005	-0.2
96	1434	38537	His3	FHEKHHSHRGY	1434.7	0.009	0.2
92	1471	251793	PRPC	GPPQGPPQGQGGPQQ	1471.7	0.003	0.1
51	1491	13325	His3	DSHAKRHHGYKR	1491.8	0.012	2.0
103	1562	43994	His3	KFHEKHHSHRGY	1562.8	0.013	-2.7
76	1680	15811	PRB1	GPRRPPQQGPGRPARPQ	1680.9	0.019	3.3
55	1718	15323	His3	KFHEKHHSHRGYR	1718.9	0.017	-1.9
86	1731	115423	PRPC	GPPQQGGHPPPPQGRPQ	1731.9	0.011	-0.1
68	1767	29397			1767.9	0.010	
88	1866	287444	PRPC	GPPPQGGRPQGPPQGQSPQ	1866.9	0.007	1.4
84	2039	68185			2040.1	0.010	
60	2077	30322			2078.1	0.016	
60	2130	43557			2131.1	0.014	
75	2178	66348			2179.1	0.011	
93	2184	37174			2185.1	0.045	
59	2240	30357			2241.2	0.014	
85	2520	81926	PRPC	GRPQGPPQQGGHQQGPPPPPQGRPQ	2521.3	0.033	-0.8
86	2624	41489			2625.4	0.064	
87	2915	144722			2916.5	0.016	
60	3000	16034			3001.5	0.200	
60	3016	27224			3017.5	0.129	
76	3017	43023			3018.5	0.134	
79	3034	45380	His3	DSHAKRHHGYKRKFHEKHHSHRGY	3035.5	0.110	5.1
67	4350	37121			4352.2	0.191	
91	4351	71239			4353.2	0.113	
106	4367	192953	PRPC	GRPQGPP....GPPQGQQSPQ	4369.2	0.017	-0.9

↑ StaggeredPMF IDs for Fig.1. Green sequences were also previously identified in the literature.

Each mass present at least 50 x in the top 40 mast list is shown, together with mapping to any of the 338 previously identified peptides from saliva. These masses will be targeted for MSMS analysis. →

Fig. 2. Saliva from Helmerhorst lab (top 8) or from me.

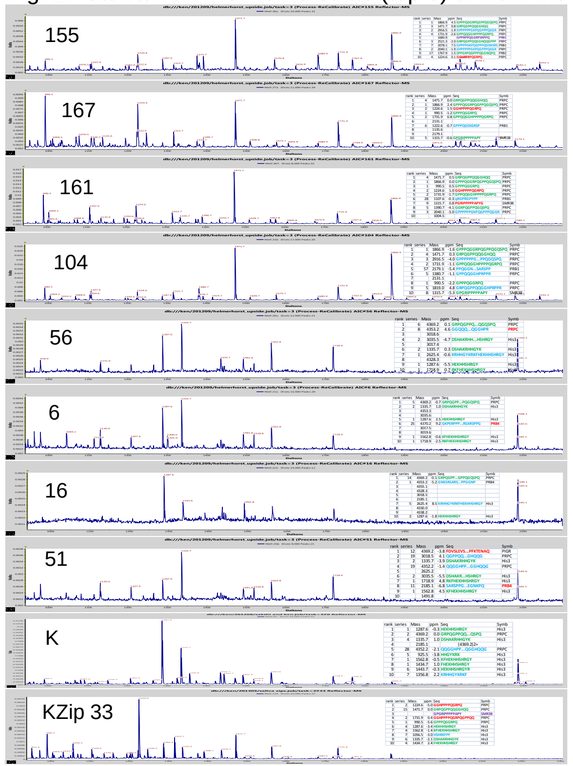
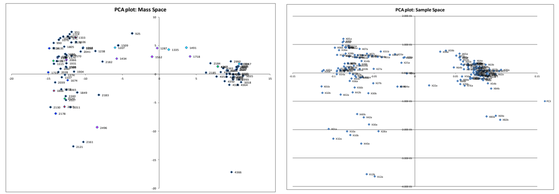


Fig.2 Legend.

The most intense 10 peaks ID'd by StaggeredPMF are listed. If green, the sequence has been published previously. If blue, the sequence is proposed. If red, the proposed sequence is different from a published sequence very similar in mass. If purple, no sequence proposed by StaggeredPMF, but appropriate peptide already published.

Fig. 3 PCA plots:

The top 40 peaks from each spectrum were used. 282 masses that were present in at least 4 samples were normalized and submitted to PCA. In the mass space plot, masses are colored according to the stagger series to which they can be mapped. Samples in which His3 stagger series are prominent map to the center of the PCA plot. Samples on the far right have a prominent 4369 peak from intact PRPC C-terminal fragment. Samples on the far left are dominated by fragments that map to PRPC stagger series. The red and dark blue masses correspond to various PRPC series, whereas the pale blue and pink series correspond to the N-ter and C-ter His3 series shown in Table 1, respectively.



Conclusions:

- Many peptides in saliva supernatants derive from series of staggered peptides with shared N or C-termini from histatins or proline-rich proteins (PRPs).
- Presumably, these derive from a combination of endopeptidases and exopeptidases.
- For PRPs, a preferred endocleavage motif is KPQ/GPP.
- Depending on subtle collection parameters, different series are most prominent.
- Can tentatively identify many peptides by high mass accuracy mapping to
  - •1.) a list of previously identified salivary peptides
  - •2.) series of peptides with shared N-termini or C-termini.
- Some of these identifications have been confirmed by MSMS.
  - •Additional identifications are in progress.
- Identifications of peptides in series is complicated by repeats, leading to multiple series with members with identical aa composition.
- PCA separates samples into sets dominated either by PRPC or His3.
- Staggered PMF may be generally useful for studying many biofluids.
- Can qualify dental hygienist according to pattern of peptides after cleaning.

References:

1.) Helmerhorst et al, Identification of KPQ as a Novel Cleavage Site Specificity of Saliva-associated Proteases.JBC 2008, 283:19957-19966.
2.) Parker KC. Scoring Methods in MALDI Peptide Mass Fingerprinting: ChemScore and the ChemApplex Program. JASMS 2002;13:22-39.