CHAPTER 24

Combining Surface Plasmon Resonance with Mass Spectrometry: Identifying Binding Partners and Characterizing Interactions

Jos Buijs* and Tohru Natsume†

**Biacore AB, Rapsgatan 7, 75450 Uppsala Sweden; †National Institute of Advanced Industrial Science and Technology. Biological Information Research Center, 2-41-6 Ohmi, Kohtoh-ku, Tokyo 135-0064, Japan*

OPTICAL BIOSENSORS HAVE COME TO BE WIDELY USED in many fields since their introduction as a research tool for the characterization of macromolecular interactions. One of the most prominent uses of biochemical sensing exploits surface plasmon resonance (SPR), a phenomenon that has been known for more than 30 years (Kretschmann 1971). For recent reviews of SPR (Winzor 2003), and its applications to proteomics, see Natsume et al. (2001).

SURFACE PLASMON RESONANCE BIOSENSORS

The central component of an SPR biosensor (as illustrated in Figure 24.1) is the sensor chip, a glass slide with a prism mounted on one side and a thin metal coating on the other side. Bait molecules are immobilized on the metal coating, most commonly in a carboxymethylated dextran film, and exposed to potential binding partners in the flow channel. A mono-

chromatic polarized light beam is directed at the sensor chip from various angles through the prism. At certain angles of incidence, the beam is totally internally reflected, and the reflected beam is detected by a sensor (the optical detection unit). When the incident beam is reflected, it produces an evanescent wave that penetrates the aqueous medium on the other side of the chip. Penetration depth is on the order of one wavelength. At a specific angle of incidence, a resonant coupling occurs between the evanescent wave and the surface plasmons. The plasmons absorb energy from the wave, causing a reduction in the intensity of the reflected beam, which can be recorded by an optical density unit. The coupling depends only on the angle of incidence and the refractive index of the medium penetrated by the evanescent wave. Changes in the density, and hence the refractive index, of the aqueous medium occur when analytes in the flow channel bind to bait molecules on the chip. This causes SPR to occur at a different angle of incidence. A sensogram, plotting resonance units against time, provides a real-time quantification of the interaction between bait and target. Application of this approach in affinity-based biosensing is discussed by Jönsson et al. (1991) and Nagata and Handa (2000). There have been several types of SPR instruments (Ward and Winzor 2000); however, this chapter describes experiments using Biacore, as it is the most commonly used instrument.

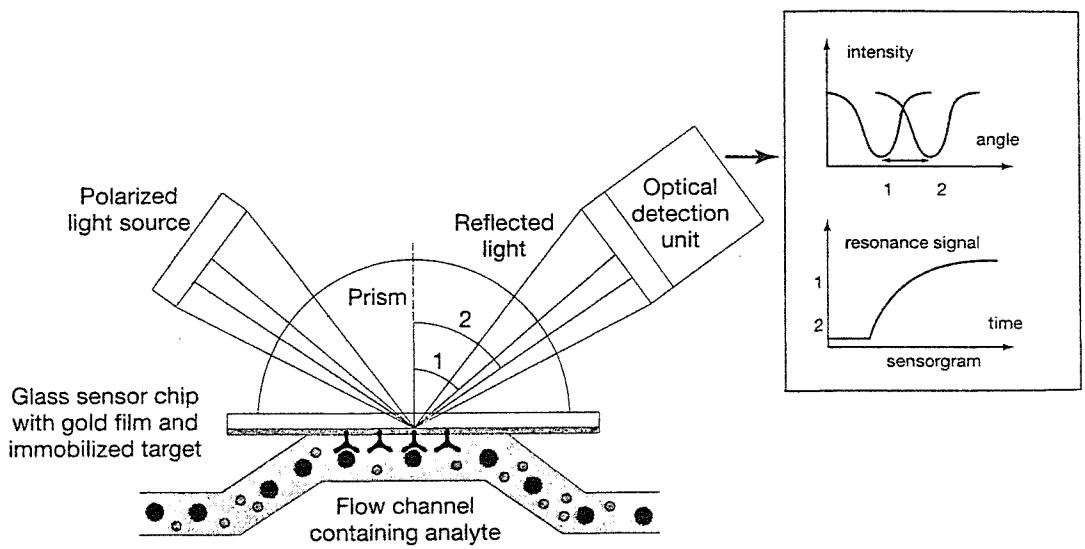

FIGURE 24.1. Schematic picture of an SPR biosensor. The polarized light is focused into a wedge-shaped light beam that illuminates the sensor chip under conditions of total internal reflection. Under these conditions, SPR results in a reduction in the intensity of light reflected from the sensor surface at a specific angle. A plot of reflected intensity against the angle shows a characteristic "dip" (*upper panel*). The angle of minimum reflectance is called the resonance angle. The resonance angle varies as a function of the refractive index of the medium near the sensor surface. A molecule (e.g., an antibody) is immobilized on the sensor chip and a sample containing a binding partner (e.g., antigen) is delivered over the sensor surface. Binding between the partners leads to an increase in the mass concentration, resulting in an increase in the refractive index and a shift in the position of the resonance angle (from 1 to 2 in the figure). The changes in the resonance angle over time are displayed in a graph called a sensorgram (*lower panel*), which is the direct representation of the interaction between the molecules on the sensor surface in real time. The unit for the SPR signal is the resonance unit (RU), where 1000 RU represents a shift in the resonance angle of 0.1°. Also note that the samples are in contact with the IFC channel in the microfluidics. To minimize the loss of the protein by adsorption to the wall of the vials or the IFC channel, use the urea-containing recovery buffer.

LIGAND FISHING AND IDENTIFICATION USING SPR BIOSENSOR AND MASS SPECTROMETRY

SPR biosensors are primarily used to characterize biomolecular interactions and are amenable to high-throughput analysis for the following reasons:

- Analysis is label-free and performed in real time, thus avoiding detection time lags.
- An auto-sample injection system allows large numbers of samples to be processed simultaneously.

The microfluidics technology incorporated into the sensor chips allows SPR to be detected in very small sample volumes (μl). SPR biosensors have been used effectively for "ligand fishing," identifying interactions between target molecules and candidate ligands (Lackmann et al. 1996; Sakano et al. 1996; Seok et al. 1997; Markgren et al. 1998; Iemura et al. 1999; Williams 2000). Unfortunately, when a binding partner is found in a biological mixture, the task of identifying the ligand is daunting. Purification of the ligand can be time-consuming and labor-intensive, often requiring the use of tailor-made strategies.

To facilitate ligand identification, SPR detection can be combined with affinity chromatography (see Chapters 10–12) and mass spectrometry (MS) (see Chapter 21 and for additional information, see Chapter 8 of Simpson 2003). MS is one of the most sensitive and specific techniques available for the identification and characterization of biomolecules. The following are two ionization mechanisms that are commonly used for MS.

- Matrix-assisted laser desorption/ionization (MALDI), in which proteins, crystallized on a sample support, are ionized by laser irradiation.
- Electrospray ionization (ESI), in which ions are generated directly from solution.

Identification of biomolecules after affinity purification with SPR biosensors has been demonstrated using both ESI-MS (Natsume et al. 2000, 2001) and MALDI-MS (Sönksen et al. 1998; Nelson and Krone 1999; Williams and Addona 2000; Catimel et al. 2001; Natsume et al. 2002). By integrating chip-based affinity purification with SPR detection, it is possible to monitor binding kinetics directly, i.e., in what proportions the molecules bind, how quickly and how tightly they bind (association and dissociation constants), and under what conditions. By combining SPR detection with MS, the molecules captured on the sensor chip can be validated or confirmed in a high-throughput manner.

Recovering samples from SPR sensor chips has been possible for many years, but one of the commercial instruments, the Biacore 3000 (Biacore), has recently been upgraded, in both software and hardware (BR-1005-75, Biacore), to enhance protein recovery (Figure 24.2). The improvements have been made in the recovery of proteins, after affinity purification, for subsequent MS analysis.

The Biacore 3000 instrument is particularly useful for ligand-fishing experiments, in which it can be used to monitor the immobilization of a bait protein, to capture the target protein and recover it from the sensor chip, and, finally, to prepare the recovered material automatically for MS analysis. These experimental steps are described in detail in the Case Study described below. A list of sensor chips suitable for the affinity purification of proteins is given in Table 24.1.

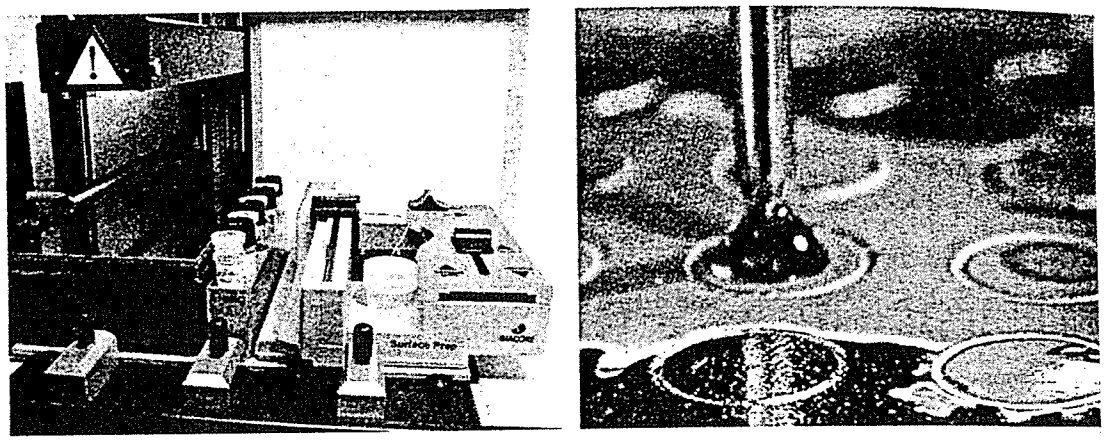

FIGURE 24.2. The Biacore 3000 for MS analysis. A 3.3x larger surface area in the Surface Prep Unit enables more efficient capturing of the bound molecules. The recovered sample is automatically delivered to a vial or deposited directly to the MALDI target located in the auto sampler area on a MALDI target holder. The enzyme solution or matrix can be deposited onto the sample on the MALDI target.

CASE STUDY: LIGAND FISHING AND IDENTIFICATION OF CALMODULIN FROM BRAIN EXTRACT

In this example, an experiment was set up to capture calmodulin from bovine brain extract on the sensor chip. The captured protein was recovered and identified using peptide mass fingerprinting. The entire process—of affinity purification, recovery of calmodulin, and sample preparation for MS analysis, including enzymatic digestion on the MALDI target—was automated using the Biacore 3000.

Calmodulin is a ubiquitous Ca^{2+}-binding protein that is found in eukaryotic cells and functions as a major physiological effector for a wide range of cellular responses. Besides the regulation of intracellular Ca^{2+} levels, calmodulin is reported to have a role in the activation of nuclear protein importation, thereby establishing a link between signal transduction and nuclear import. Myosin light-chain kinase is one of the physiological interaction partners of calmodulin. A peptide derived from this protein, containing the calmodulin-binding domain, was used as bait for calmodulin.

The protocol used is detailed below with the experimental steps, as controlled by the Biacore 3000 software, summarized in Table 24.2. All experiments were performed at 25°C and the water used was obtained from a Milli-Q water purification system (Millipore).

TABLE 24.1. Sensor chips that can be used for affinity purification using the Biacore 3000 biosensor

Sensor Chip[a]	Surface Coating	Applications
CM5 (BR-1000-14)	carboxymethylated dextran	general purpose
NTA (BR-1000-34)	carboxymethylated dextran, preimmobilized with NTA	capture of histidine-tagged biomolecules
SA (BR-1000-32)	carboxymethylated dextran, preimmobilized with streptavidin	captures biotinylated biomolecules
CM4 (BR-1005-39)	carboxymethylated dextran with lower degree of carboxylation than CM5	reduces nonspecific binding of highly positively charged molecules as may be found in cell lysates

[a]Information in parentheses indicates Biacore catalog numbers.

TABLE 24.2. Sequential listing of the experimental steps for a ligand fishing experiment, including the utilized Biacore 3000 commands

Process	Command	Volumes, Solution, and Flow Rates
Equilibrate system with buffer	prime	HBS-EP
Surface activation	mix inject	100 µl of EDC with 100 µl of NHS 35 µl of EDC/NHS, 5 µl/min
CBD immobilization	inject	50 µl of CBD, 5 µl/min
Deactivation	inject	35 µl of EtAm, 5 µl/min
Condition	quickinject	10 times 5 µl of 50 mM NaOH, 5 µl/min
Change buffer	prime	3 times HBS-N
Injection of bovine brain extract	MS_inject	200 µl of brain extract, 20 µl/min, 200 µl of MS buffer
Wash	MS_wash	30 µl of NaOH, 40 µl of NaOH, 200 µl of MS buffer
Recovery	MS_recover	50 µl of recovery buffer, 250 µl of H_2O, 300 µl of MS buffer
Add	microtransfer	1 µl of trypsin
Digestion	wait 900	wait for 15 minutes
Add	microtransfer	1 µl of TFA
Add	microtransfer	1 µl of matrix solution

The type of solutions, their volumes, and if applicable, the flow rate at which solution is injected in the instrument are given in column 3.

MATERIALS

CAUTION: See Appendix 7 for appropriate handling of materials marked with <!>.

◗ Reagents

Calmodulin-binding domain (CBD) (0.1 mg) (CalBiochem)

EDC (1-ethyl-3-[3-dimethylaminopropyl] carbodiimide hydrochloride) (e.g., Sigma or Biacore) <!>

Ethanolamine-HCl (1 M, pH 9.0) (e.g., Sigma or Biacore) <!>

HBS-EP

- 10 mM HEPES (pH 7.4)
- 150 mM NaCl
- 3 mM EDTA
- 0.005% Surfactant P20

Surfactant P20 is polyoxyethylenesorbitan, a nonionic surfactant recommended for inclusion in the buffers used in Biacore systems. It is available from Biacore, where it has been tested for peroxides and carbonyls, and is supplied as a sterile-filtered 10% solution in water.

HBS-N: 10 mM HEPES containing 150 mM NaCl (pH 7.4)

HBS-N containing 2 mM $CaCl_2$ <!>

Initial buffer for microcolumn: 0.1% formic acid (100 µl required) <!>

Matrix solution

Dissolve 0.5 mg of α-cyano-4-hydroxycinnamic acid <!> (Bruker) in 670 µl of ethanol <!> and 330 µl of acetone <!>.

MS buffer: 50 mM ammonium bicarbonate <!> and 2 mM $CaCl_2$ (pH 8.2)
For MS-compatible buffers, see note in Stage 2.

NaOH (50 mM) <!>

N-hydroxysuccinimide (e.g., Sigma or Biacore) <!>
Aliquots (100 µl) of *N*-hydroxysuccinimide should be stored at –20°C.

Potassium phosphate buffer (10 mM, pH 7.4) <!>

Recovery solution: 0.5 mM EGTA and 50 mM NH_4HCO_3 (pH 8.2)
For recovery solutions, see note in Stage 3.

Running buffer

Trifluoroacetic acid (TFA) (0.5% v/v) <!>

Trypsin <!>, Sequence-grade modified (Promega), 15 µg/ml in 60% acetonitrile <!>

Wash solutions
For suitable wash solutions, see Step 9.

◗ Equipment

Biacore 3000 instrument equipped with sample rack, autosampler, and software (version 4.0 or higher).
For information on analyte recovery using an earlier version of the Biacore 3000, contact Biacore directly (www.biacore.com).

CM5 sensor chip (Biacore)

MALDI target (e.g., Scout MTP, Bruker, ABI Voyager, MALDI-Qstar, Micromass Qtof)
If using the alternative procedure for Stage 4, the analyte is collected in a vial rather than being deposited on a MALDI target.

Mass spectrometer (e.g., Autoflex MALDI-TOF from Bruker)

◗ Biological Sample

Bovine brain extract: 10 mg of bovine brain acetone powder (Sigma)

◗ Additional Materials

The alternative Stage 4 of this protocol requires the following additional reagents:
Microcolumn 1 elution buffer: 30% acetonitrile <!>, 0.1% formic acid <!>
Microcolumn 2 elution buffer: 20 mg/ml sinapinic acid <!>, 60% acetonitrile, 0.1% trifluoroacetic acid <!>

For ESI-MS analysis
ESI or tandem mass spectrometer (ESI-LC-MS or ESI-LC-MS/MS)
Nanospray tip (e.g., nanoflow probe tip [Micromass]; nanobore stainless steel emitters and stage tips [Proxeon Biosystems A/S]; and PicoTips [New Objective, Inc.])

Formic acid (0.1% v/v) in H_2O

Incubator set at 37°C

For MALDI analysis
MALDI target
MALDI-TOF mass spectrometer
Sinapinic acid crystals (20 mg/ml sinapinic acid in acetone)

Micro-reversed-phase POROS column prepared in a GELoader tip according to the method of Gobom et al. (1999)

Trifluoroacetic acid (TFA) (50 nl)

Trypsin digestion buffer (pH 9.0): 100 mM Tris <!>, 4 M urea <!>, and 0.01% *n*-octylglucopyranoside
Trypsin (15 μg/ml) in trypsin digestion buffer
Trypsin, sequencing-grade (Promega)

METHOD

Stage 1: Immobilization of the CBD to a CM5 Sensor Chip

The CBD was immobilized to a carboxymethylated dextran of the sensor chip by amine coupling chemistry (see Johnsson et al. 1991). In this procedure, the molecules are covalently coupled via their primary amine groups to the sensor chip. The coupling provides the high degree of immobilization necessary for successful recovery experiments. The number of molecules captured on the chip is directly proportional to the degree of immobilization and the number of flow cells. In a Biacore 3000, four flow cells, each with a surface area of 1.2 mm^2, are available for this purpose.

1. Equilibrate the system with HBS-EP using the PRIME command.
2. Thaw 100-μl aliquots of EDC and NHS, vortex, and place them in the appropriate slots in the sample rack. Activate the sensor chip surface with 200 μl of a 1:1 mixture of these solutions using the MIX command.
3. Prepare a solution of CBD by dissolving 0.1 mg of CBD in 1 ml of 10 mM potassium phosphate buffer (pH 7.4) and divide it into 100-μl aliquots. INJECT 35 μl of EDC/NHS at a flow rate of 5 μl/min. Then INJECT 50 μl of CBD, at the same rate, over the activated sensor surface for coupling.
4. After coupling the CBD, deactivate excess active groups on the sensor surface by injecting 35 μl of ethanolamine at a flow rate of 5 μl/min. To condition the sensor, QUICKINJECT ten 5-μl aliquots of 50 mM NaOH at 5 μl/min.

Stage 2: Capturing Calmodulin (CaM) on the Sensor Chip

5. Prepare the bovine brain sample from bovine brain acetone powder by suspending 10 mg in 1 ml of HBS-N containing 2 mM $CaCl_2$. Centrifuge the extract at 16,000*g* (Biofuge pico, Heraeus Instr.) and discard the insoluble pellet. Dilute the supernatant 10 times in the same buffer.
6. After ~1 minute of running buffer flow, PRIME the system with three applications of HBS-N.

   The HBS-N is a low-salt MS-compatible buffer, which replaces the higher-salt running buffer (see the shaded panel below regarding MS-compatible buffers), as can be seen from the dip (at ~50 seconds) in the sensorgram shown in Figure 24.3. The dip occurs because the lower-salt buffer has a lower refractive index than the running buffer (see the section on Surface Plasmon Resonance Biosensors).

7. MS_INJECT 200 μl of brain extract (20 μl/min), followed by 200 μl of MS buffer over the CM5 chip immobilized with CBD.

   Generally, flow rates ranging from 1 to 20 μl/min and sample volumes on the order of a few tenths to hundreds of microliters are used to capture the target molecule.

8. Directly after sample injection, flush the integrated fluidic system (IFC), including the flow cell area, with an MS-compatible buffer (HBS-N) for 6 seconds.

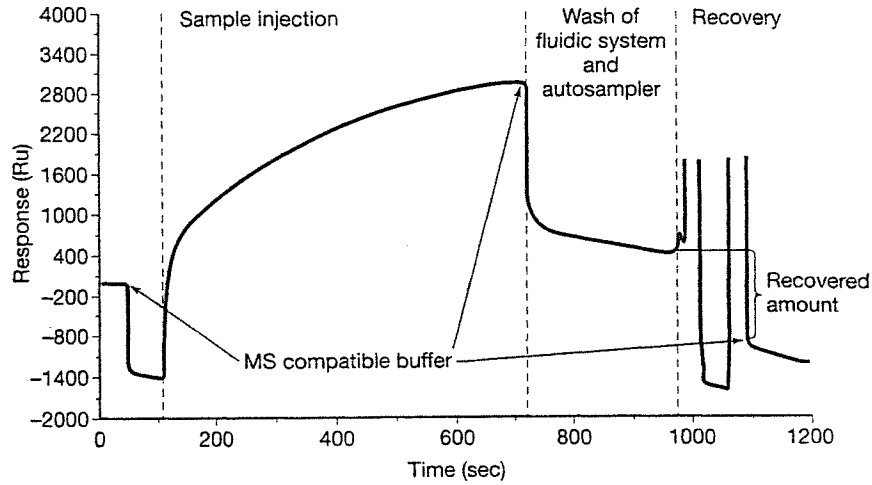

FIGURE 24.3. A typical sensorgram, showing the SPR signal (in response units, RU) during an affinity purification and recovery experiment. A signal of 1 RU in one flow cell corresponds to 1 pg of protein (see Stenberg et al. 1991).

MS-COMPATIBLE BUFFER

MS-compatible buffer is used throughout the affinity purification experiment. It is crucial for the MS sensitivity of the recovered sample and has three functions.

- It washes the flow cells when the target protein is captured on the chip.
- It separates carryover contamination from chemicals present in the running buffer and in the injected sample.
- It is used before and after the various steps in the affinity purification and recovery experiment to quantify bound and recovered levels from the sensorgram.

The MS-compatible buffer provides a suitable environment to maintain the integrity of the biomolecular interaction, but it contains as little salt as possible so that it does not interfere with the MS analysis. Volatile buffers such as ammonium bicarbonate are compatible with MS analysis as they evaporate during the MALDI sample preparation step. Buffers containing organic acids in combination with ammonium, such as ammonium acetate, citrate, or formate, are also compatible when used at lower concentrations. More common buffers can affect MS sensitivity in various ways. For a detailed list of the compatibility of buffers with MALDI-MS analysis, see Kallweit et al. (1996). Phosphate, for example, is known to have a strong negative effect, whereas other buffers, such as Tris and HEPES, are more compatible with MALDI-MS. The MS-compatible buffer in this experiment contains 50 mM ammonium bicarbonate and 2 mM $CaCl_2$. The $CaCl_2$ was added because Ca^{2+} is required for the binding of calmodulin to CBD.

By using a desalting micro-reversed-phase column (see below An Alternative Method for Stage 4: Sample Preparation for MS Analysis), even buffers that are incompatible with MS ionization can be used (for further information on desalting micro-reversed-phase columns, see Chapter 8 [including Protocol 2] of Simpson 2003). A chaotropic reagent such as urea can be added to denature recovered sample prior to digestion. (Some proteins can be digested only under denaturing conditions.) Urea-containing recovery buffer can also be used to minimize protein adsorption to the wall of the vial or the IFC channel (see Figure 24.1 and its legend), leading to more efficient recovery of low-abundance or sticky proteins.

Stage 3: Recovery of Calmodulin from the Sensor Chip

9. Wash the autosampler and IFC, but not the flow cells, sequentially with three solutions using the MS_WASH command, as follows.
   - Use a mild basic or acidic solution (e.g., 50 mM NaOH, 1–5% TFA, formic acid, or acetic acid) or low-concentration organic solvents (10–20% acetonitrile or methanol) for the first two wash solutions.
   - For the third wash, use an MS-compatible buffer to equilibrate the IFC prior to sample recovery.
10. Recover the bound calmodulin using the MS_RECOVER command, which applies 2 µl of recovery solution to the sensor chip, and incubate it for 30 seconds.
11. Draw the 2-µl portion back into the autosampler needle. For peptide fingerprinting, deposit the sample on a Scout MTP MALDI target (Bruker) with a spot size of 0.8 mm and continue with Stage 4 below. For material recovered in non-MS-compatible buffer, collect the sample in a vial and continue with the alternative Stage 4 procedure.

During the MS_RECOVER process, air is passed over the flow cells for 30 seconds before and after the recovery solution is incubated in the flow cells. When air is flowing through the flow cells, the SPR signal is temporarily disrupted, as seen in the sensogram (Figure 24.3).

Recovery Solutions

The selection of the recovery solution is directly related to the way in which the sample is prepared for MS analysis. MS analysis often requires acidification of the sample to generate abundant positively charged analytes. At the same time, low concentrations of organic acids, e.g., 0.5% TFA, are often suitable to release the captured target molecules from the sensor chip.

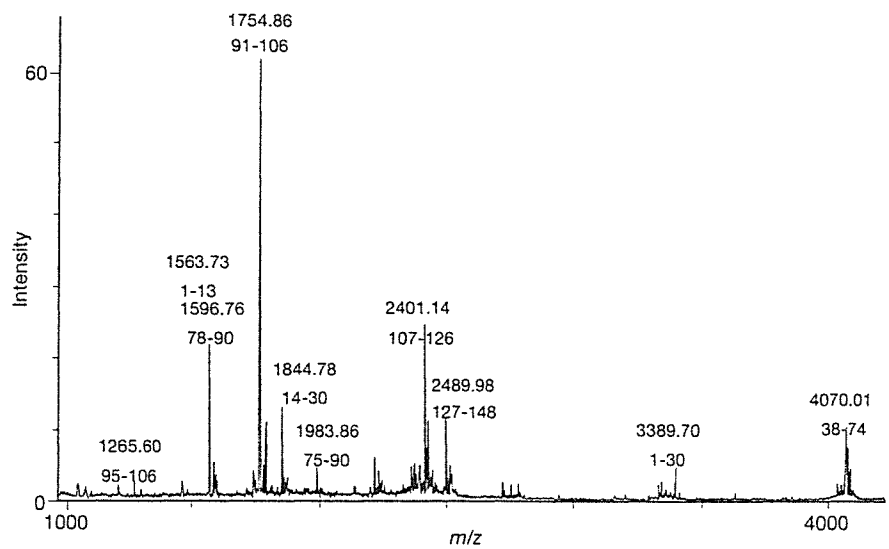

FIGURE 24.4. MALDI spectrum of a tryptic digest of the recovered target protein after injection of brain extract over a sensor chip, modified with the calmodulin-binding domain. The *m/z* values of these mass peaks were submitted to a database (Mascot; for details on searching databases using MS data, see Search Engines for Identifying Proteins Using MS Data in Chapter 8 of Simpson [2003]) and bovine calmodulin was unambiguously identified. The sequence coverage was 95%. Three peptide ions (1563.73, 2401.14, and 3389.70 *m/z*) were not identified in the first database search. Two additional peptides (1563.73 and 2401.14 *m/z*) were assigned when the posttranslational modifications were incorporated in the database correlation (acetylated amino terminus and trimethylated Lys-114).

Stage 4: Preparation of the Recovered Material for Mass Spectrometry Analysis

12. For peptide mass fingerprinting, digest the target protein on the MALDI plate with trypsin prior to adding the matrix. Use the MICROTRANSFER command to add 1 µl of trypsin (15 µg/ml in 60% acetonitrile) to the recovered material (final concentration 5 µg/ml) on the MALDI target. Allow it to dry (~15 minutes, use the WAIT 900 command).

    For additional details on preparing samples for peptide mass fingerprinting, see Chapters 8 and 9 (including Protocol 3) in Simpson (2003).

13. After digestion, acidify the sample using the MICROTRANSFER command to add 1 µl of TFA (0.5% v/v) and then 1 µl of a matrix solution.

14. Record mass spectra using a mass spectrometer in the reflection mode with delayed extraction, and submit the *m/z* values of the mass peaks to a database.

    A typical mass spectrum and database search is shown in Figure 24.4. For details on searching databases using MS data, see Search Engines for Identifying Proteins Using MS Data in Chapter 8 of Simpson (2003).

IMPORTANT

- As matrix solutions tend to leave crystals on the autosampler needle, an additional wash of the needle in organic solvents may be necessary.
- General guidelines for MALDI sample preparations, including digestion, are described by Kussmann and Roepstorff (2000). Note that the ionization process in MS measurements is extremely sensitive to contaminants such as detergents (SDS, Tween), buffer salts, and even leakage of chemical substances from column material and vials. In cases where these problems occur, samples should be cleaned up using a micro-reversed-phase column (see below An Alternative Method for Stage 4: Sample Preparation for MS Analysis; also for further information on desalting micro-reversed-phase columns, see Chapter 8 [including Protocol 2] in Simpson 2003).

An Alternative Method for Stage 4: Sample Preparation for MS Analysis

When samples are recovered in a vial, instead of being deposited on a MALDI target, it is important to minimize the delay between their recovery and their preparation for MS. Proteins stick to almost all surfaces with which they come in contact. Because of the small volumes and low concentrations used, losses through adsorption and sample handling can be significant. Adsorption can be avoided by adding urea and *n*-octylglucopyranoside (an MS-compatible detergent) to the recovery buffer. If this proves to be necessary, samples must also be cleaned up and desalted prior to MS analysis. The $ZipTip_{micro\text{-}C18}$ (Millipore) is a column suitable for cleaning up, concentrating, and desalting samples; samples can be eluted in as little as 0.5 µl of solution. Even smaller elution volumes (down to 50–100 nl) are possible using a homemade POROS reversed-phase (ABI) microcolumn constructed inside a GELoader tip (Eppendorf) (see Gobom et al. 1999; Simpson 2003). Samples recovered from these columns in non-MS-compatible buffer can be analyzed using MALDI or ESI-MS. The following procedure describes the preparation of material recovered in a non-MS-compatible buffer.

METHOD

1. Digest the target protein with trypsin by adding 1 µl of 15 µg/ml trypsin in trypsin digestion buffer to the vial. Incubate for 2 hours at 37°C.
2. Desalt the digestion mixture and concentrate the sample using a micro-reversed-phase POROS column. Equilibrate the column with 0.1% formic acid (FA) in H_2O (initial buffer), and pass the trypsin digestion mixture over the column. Then wash the column with the same initial buffer.

   IMPORTANT: Due to the high back pressure, the micro-reversed-phase POROS packing is the only one recommended.

3. Complete one of the two following procedures.

   For ESI-MS analysis:

   a. Elute the peptide mixture in 0.5–1 µl of Microcolumn 1 elution buffer.

   b. Load the sample into the capillary using either a fused silica syringe needle or a gel loader tip. Apply pressure to the back of the tip until a drop of liquid is seen.

   c. Tune the source voltage, and adjust the gas flow for maximum ion current. Record mass spectra. For details of the operation, see Wilm and Mann (1996).

   For MALDI analysis:

   a. Elute the peptide mixture in 50 nl of Microcolumn 2 elution buffer, directly onto a nano-spray tip.

   b. Cover the deposited sample with a thin layer of sinapinic acid crystals (20 mg/ml sinapinic acid in acetone).

   c. Record mass spectra using a MALDI-TOF mass spectrometer according to the manufacturer's instructions.

ACKNOWLEDGMENTS

We thank A. Zhukov and Ö. Jansson for their contributions in developing the recovery functionality in the Biacore 3000 and for experimental data, and S. Hashimoto for editing the manuscript.

REFERENCES

Catimel B., Rothacker J., and Nice E. 2001. The use of biosensors for microaffinity purification: An integrated approach to proteomics. *J. Biochem. Biophys. Methods* **49:** 289–312.

Gobom J., Nordhoff E., Mirgorodskaya E., Ekman R., and Roepstorff P. 1999. Sample purification and preparation technique based on nano-scale reversed-phase columns for the sensitive analysis of complex peptide mixtures by matrix-assisted laser desorption/ionization mass spectrometry. *J. Mass Spectrom.* **34:** 105–116.

Iemura S., Yamamoto T.S., Takagi C, Kobayashi H., and Ueno N. 1999. Isolation and characterization of bone morphogenetic protein-binding proteins from the early *Xenopus* embryo. *J. Biol. Chem.* **274:** 26843–26849.

Johnsson B., Löfås S., and Lindquist G. 1991. Immobilization of proteins to a carboxymethyldextran modified gold surface for biospecific interaction analysis in surface plasmon resonance. *Anal. Biochem.* **198:** 268–277.

Jönsson U., Fagerstam L., Ivarsson B., Johnsson B., Karlsson R., Lundh K., Löås S., Persson B., Roos H., and Rönnberg I., Sjölander S., Stenberg E., Ståhlberg R., Urbaniczky S., Östlin H., and Malmqvist M. 1991. Real-time biospecific interaction analysis using surface plasmon resonance and a sensor chip technology. *BioTechniques* **11:** 620–627.

Kallweit U., Börnsen K.O., Kresbach G.M., and Widmer H.M. 1996. Matrix compatible buffers for analysis of proteins with matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **10:** 845–849.

Kretschmann E. 1971. Die Bestimmung optischer Konstanten von Metallen durch Anregung von Oberflächenplasmaschwingungen. *Z. Physik* **241:** 313–324.

Kussmann M. and Roepstorff P. 2000. Sample preparation techniques for peptides and proteins analyzed by MALDI-MS. *Methods Mol. Biol.* **146:** 405–424.

Lackmann M., Bucci T., Mann R.J., Kravets L.A., Viney E., Smith F., Moritz R.L., Carter W., Simpson R.J., Nicola N.A., Mackwell K., Nice E.C., Wilks A.F., and Boyd A.W. 1996. Purification of a ligand for the EPH-like receptor HEK using a biosensor-based affinity detection approach. *Proc. Natl. Acad. Sci.* **93:** 2523–2527.

Markgren P.O., Hamalainen M., and Danielson U.H. 1998. Screening of compounds interacting with HIV-1 proteinase using optical biosensor technology. *Anal. Biochem.* **265:** 340–350.

Nagata K. and Handa H., eds. 2000. *Real-time analysis of biomolecular interactions: Applications of BIACORE.* Springer-Verlag, Tokyo, Japan.

Natsume T., Nakayama H., and Isobe T. 2001. BIA-MS-MS: Biomolecular interaction analysis for functional proteomics. *Trends Biotechnol.* **19** s28–s33.

Natsume T., Nakayama H., Jansson Ö., Isobe T., Takio K., and Mikoshiba K. 2000. Combination of biomolecular interaction analysis and mass spectrometric amino acid sequencing. *Anal. Chem.* **72:** 4193–4198.

Natsume T., Taoka M., Manki H., Kume S., Isobe T., and Mikoshiba K. 2002. Rapid analysis of protein interactions: On-chip micropurification of recombinant protein expressed in *Escherichia coli*. *Proteomics* **2:** 1247–1453.

Nelson R.W. and Krone J.R. 1999. Advances in surface plasmon resonance biomolecular interaction analysis mass spectrometry (BIA/MS). *J. Mol. Recog.* **12:** 77–93.

Sakano S., Serizawa R., Inada T., Iwama A., Itoh A., Kato C., Shimizu Y., Shinkai F., Shimizu R., Kondo S., Ohno M., and Suda T. 1996. Characterization of a ligand for receptor protein-tyrosine kinase HTK expressed in immature hematopoietic cells. *Oncogene* **13:** 813–822.

Seok Y.J., Sondej M., Badawi P., Lewis M.S., Briggs M.C., Jaffe H., and Peterkofsky A. 1997. High affinity binding and allosteric regulation of *Escherichia coli* glycogen phosphorylase by the histidine phosphocarrier protein, HPr. *J. Biol. Chem.* **272:** 26511–26521.

Simpson R.J. 2003. *Proteins and proteomics: A laboratory manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Sönksen C.P., Nordhoff E., Jansson Ö., Malmqvist M., and Roepstorff P. 1998. Combining MALDI mass spectrometry and biomolecular interaction analysis using a biomolecular interaction analysis instrument. *Anal. Chem.* **70:** 2731–2736.

Stenberg E., Persson B., Roos H., and Urbaniczky C. 1991. Quantitative determination of surface concentration of protein with surface plasmon resonance by using radiolabeled proteins. *J. Colloid Interface Sci.* **143:** 513–526.

Ward L.D. and Winzor D.J. 2000. Relative merits of optical biosensors based on flow-cell and cuvette designs. *Anal. Biochem.* **285:** 179–193.

Williams C. 2000. Biotechnology match making: Screening orphan ligands and receptors. *Curr. Opin. Biotechnol.* **11:** 42–46.

Williams C. and Addona T.A. 2000. The integration of SPR biosensors with mass spectrometery: Possible applications for proteome analysis. *Trends Biotechnol.* **18:** 45–48.

Wilm M. and Mann M. 1996. Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **68:** 1–8.

Winzor D.J. 2003. Surface plasmon resonance as a probe of protein isomerization. *Anal. Biochem.* **318:** 1–12.

Gene expression profile in rat pancreatic islet and RINm5F cells

H Wang[1], Y Horikawa[1,2,3], L Jin[1], T Narita[1], S Yamada[1], N Shihara[1,2], K Tatemoto[4], M Muramatsu[3], T Mune[3] and J Takeda[1,2,3]

[1]Laboratory of Molecular Genetics, Department of Cell Biology, Institute for Molecular and Cellular Regulation, Gunma University, Gunma, Japan

[2]Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Corporation (JST), Kawaguchi, Japan

[3]Department of Diabetes and Endocrinology, Division of Molecule and Structure, Gifu University School of Medicine, Gifu, Japan

[4]Laboratory of Peptide and Protein Research, Department of Molecular Physiology, Institute for Molecular and Cellular Regulation, Gunma University, Gunma, Japan

(Requests for offprints should be addressed to Yukio Horikawa, Department of Diabetes and Endocrinology, Gifu University School of Medicine, 1-1 Yanagido, Gifu-city, Gifu 500-1194, Japan; Email: yhorikaw@cc.gifu-u.ac.jp)

Abstract

To clarify tissue-specificity of pancreatic β cells, comparison of mRNA expression in various conditions of the tissue of multiple organisms is important. Although the developed methodologies for mRNA monitoring such as microarray, rely on the growth of dbEST (database of expressed sequence tag), a large number of unknown genes in the genome, especially in the rat, have not been shown to be expressed. In this study, we have established the first database of ESTs from rat pancreatic islet and RINm5F cells. Two cDNA libraries were constructed using mRNAs from rat pancreatic islet and RINm5F cells to cover a wider spectrum of expressed genes. Over 40 000 clones were randomly selected from the two libraries and partially sequenced. The sequences obtained were subjected to BLAST database analyses. This large-scale sequencing generated 40 710 3′-ESTs. Clustering analysis and homology search of nucleotide and peptide databases using both 3′- and 5′-ESTs revealed 10 406 non-redundant transcripts representing 4078 known genes or homologs and 6328 unknown genes. To confirm actual expression, the unknown sequences were further subjected to dbEST search, resulting in the identification of 5432 significant matches to those from other sources. Interestingly, of the remaining sequences showing no match, 779 were found to be encoded by exon–intron organization in the corresponding genomic sequences, suggesting that these are newly found as actually expressed in this study. Since many genes are up- or down-regulated in differing conditions, applications of the expression profile should facilitate identification of the genes involved in cell-specific functions in normal and disease states.

Journal of Molecular Endocrinology (2005) **35,** 1–12

Introduction

Pancreatic islets play the critical role in the regulation of blood glucose by secreting hormones from endocrine cells that differentiate from common progenitor cells during fetal development (postnatal origin of β-cell replenishment remains controversial) (Edlund 2002, Bonner-Weir & Sharma 2002, Dor *et al.* 2004). Functional defects of β-cells lead to the development of diabetes mellitus. Since a number of genes are involved in the pathogenesis of impaired insulin secretion, it is important to characterize the expression profile of a set of genes that endow β-cells with the tissue-specific functions of insulin synthesis and secretion.

Recent genome projects demonstrated a similar number of 30 000–40 000 genes in human and mouse chromosomes, including ~27 000 protein-encoding transcripts for which there was strong corroborating evidence and ~10 000 computationally derived genes with weak supporting evidence (International Human Genome Sequencing Consortium 2001, Venter *et al.* 2001, Mouse Genome Sequencing Consortium 2002). Comparison with the transcriptome revealed almost all of the human genes known to be expressed to have orthologues in the mouse genome. The other putatively novel genes in the genome were detected using computer algorithms for transcript prediction. To estimate the accuracy of the power of new gene detection, the results of the gene annotation done by the two human genome efforts were previously compared (Hogenesch *et al.* 2001). Surprisingly, although a similar number of the total genes was demonstrated, there is little agreement regarding the new genes predicted by the two projects, suggesting that a significant fraction of tissue-restricted transcripts for novel genes remain undiscovered, possibly due to limitations in the computer prediction methods.

As expression analysis of the genes in multiple organisms becomes a major focus in the new era of biology, functional genomics will rely largely on the vast

0952–5041/05/035–001
DOI: 10.1677/jme.1.01738
Online version via http://www.endocrinology-journals.org

sources of subsets of partial cDNA sequences from various tissues that have proven enormously valuable and are deposited as expressed sequence tags (ESTs) in the public databases. The Endocrine Pancreas Consortium has recently constructed human and mouse cDNA libraries from various conditions of endocrine pancreas and generated over 100 000 ESTs (Bernal-Mizrachi *et al.* 2003). We also have collected ~20 000 ESTs from human normal pancreatic islets and islet tumors, resulting in the identification of ~3000 new genes expressed in the islets (Takeda *et al.* 1993, Jin *et al.* 2003). Such systematic sequencing efforts complement each other and should improve the various methodologies including DNA microarray technology (Scearce *et al.* 2002) for monitoring differential gene expression in normal and disease states. In addition, the laboratory rat is an indispensable model organism of human diseases, providing a useful tool in experimental medicine and drug discovery. As various spontaneous diabetic rats such as the GK and OLETF rats and the experimental streptozotocin-induced diabetic rat are widely used in pancreatic islet studies, it is important to establish an additional source of rat expressed sequences. However, although ~26 millions of ESTs have so far been deposited in the database (dbEST release 031105), approximately 40% of which are derived from human and mouse, only 2·6% of the sequences are from rat, and none are from pancreatic islets except for the present deposition. In this study, toward elucidation of the entire transcriptome in rat pancreatic islets, we have made two cDNA libraries, one from rat normal pancreatic islets and the other from RINm5F tumor cells having undergone less differentiation (Gazdar *et al.* 1980, Philippe *et al.* 1987), and performed a large-scale collection of ESTs. Since a number of the genes are up- or down-regulated in different conditions, a collection of ESTs from these distinct cDNA sources should more effectively cover a wider spectrum of expressed genes, generating a larger pool of non-redundant sequences. In addition, since the insulin content of the less-differentiated RINm5F cells used in this study was previously found to be much lower than that of normal islets (Kayo *et al.* 1997), direct comparison of the expression profiles of the two cDNA libraries should facilitate the identification of the genes involved in insulin synthesis and secretion as well as in β-cell differentiation and tumorigenesis.

Materials and methods

Preparation of rat pancreatic islets

Pancreatic islets were prepared from male Wister rats by a collagenase digestion method as described previously (Ma *et al.* 1996). Briefly, under pentobarbital anesthesia, the pancreas was distended by an injection of 10 ml Hank's solution containing 0·3 mg/ml collagenase (type XI, Sigma-Aldrich, StLouis, MO, USA). Islets were separated by the Ficoll (Amersham, Piscataway, NJ, USA) density gradient method with four layers (27%, 23%, 20·5%, and 11% of Ficoll dissolved in Hank's solution). After centrifugation at 450 ***g*** for 15 min, pancreatic islets were concentrated at the interface between the 11% and 20·5% Ficoll layers. Islets were then harvested by a pick-up method under a stereomicroscope. Purity of the islets was estimated to be ~99% by counting the cells immunoreactive to insulin and glucagon antibodies after trypsin treatment of a fraction of the islets collected. The high purity also could be estimated by the frequency of cDNA for major exocrine molecules, such as α-amylase, in the entire islet ESTs identified, as described below.

Large-scale cDNA sequencing

Two unidirectional cDNA libraries were constructed in the Uni-ZAP XR vector (Stratagene, La Jolla, CA, USA) using mRNAs from rat normal pancreatic islets and RINm5F cells. A large set of plasmid DNAs for sequencing was prepared as described (Takeda *et al.* 1993, Jin *et al.* 2003). Briefly, the cDNA libraries were excised *en masse* from the λ phage into phagemid particles using the ExAssist phage system (Stratagene), and subsequently transfected into *E. coli* SOLR (Stratagene) for conversion to plasmid forms. Plasmid DNAs were extracted from *E. coli* colonies randomly selected from LB-Amp plates using the Biomek 2000 mini-prep system (Beckman, Fullerton, CA, USA). Single-pass DNA sequencing from the 3′-end of the inserts was performed using a BigDye Terminator Cycle Sequencing FS Ready Reaction Kit and DNA Sequencer model 3700 (Applied Biosystems, Foster City, CA, USA). Vector sequences were removed from the results using Assembly LIGN software (Oxford Molecular Group PLC, Oxford, UK). Quality assessment of the sequences obtained was performed using PE Sequencing Analysis 3·3 software (Applied Biosystems).

Database analysis of rat pancreatic islet and RINm5F ESTs

We compared a total of ~40 000 sequences from rat pancreatic islet and RINm5F cells with non-redundant nucleotide and peptide sequences extracted *in silico* from databases at the National Center for Biotechnology Information (NCBI). Before comparisons, interspersed repetitive sequences such as LINEs were unmasked and removed from the pool using software RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html). To assemble sequences sharing a stretch of nucleotide identity, the LaboServer system (World fusion, Tokyo, Japan) was applied to make contigs.

Representative sequences from each contig then were subjected to BLASTN analysis for sequence homology at nucleotide level against a merged database by the Kiroku program (World fusion). If a query sequence shared over 95% nucleotide identity and showed a score of more than 400 with any sequences in the database, they were grouped together. The clones without significant match to known sequences in the nucleotide database were re-sequenced from the other end to compare the sequences with those in the peptide database at NCBI using BLASTX program (Altschul *et al.* 1997), which conceptually translates the query sequence in all six reading frames for comparison. The ESTs identical or highly similar to functionally annotated genes were first classified into seven major categories according to the general functions of the proteins encoded, and then further classified into subcategories according to their specific functions.

Semi-quantitative RNA expression analysis

To ascertain the level of mRNA expression of the ESTs from rat normal islets and RINm-5F cells, real-time quantitative reverse transcriptase (RT)-PCR was carried out using an ABI PRISM 7900HT Sequence Detection System (Applied Biosystem). Total RNA was extracted from pooled islets isolated from normal rats and RINm5F cells, using an RNeasy Mini Preparation Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. TaqMan primers and probes were designed using Primer Express software purchased from Applied Biosystems. TaqMan reactions were performed in a reaction volume of 20 µl using components supplied in a TaqMan PCR reagent kit. Each reaction consisted of 10 µl TaqMan Universal Master Mix, 900 nM of each amplification primer, and 250 nM corresponding TaqMan probe. Each sample was run for an initial 2 min at 50 °C and 10 min at 95 °C, followed by 40 cycles at 95 °C for 15 s and at 60 °C for 1 min. Amplification data were collected by the 7900HT Sequence Detector and analyzed using Sequence Detection System software. The RNA concentration was determined from the threshold cycle at which fluorescence is first detected, cycle number being inversely related to RNA concentration.

In situ hybridization

ESTs showing marked differences in frequency between islet and RINm5F cells were subjected to analysis of mRNA distribution in the pancreas by *in situ* hybridization. Paraffin embedded blocks and sections of normal rat pancreas for *in situ* hybridization (ISH) were obtained from GENOSTAFF, Inc. (Tokyo, Japan). The pancreases of male Wistar rats 8 weeks old (CREA Tokyo, Japan, Inc.) were dissected after perfusion, fixed by Tissue Fixative (GENOSTAFF, Cat No.STF-01), and embedded in paraffin by the proprietary procedures.

ISH was performed with the Ventana HX system (Ventana Medical Systems, Inc., Tucson, AZ, USA). Entire EST inserts were amplified by PCR using ExTaq (TaKaRa, Kyoto, Japan) in a 50 µl reaction mixture using M13 forward and reverse primers. Amplification was performed as follows: 3 min at 94 °C for initial denaturation, 35 cycles of 94 °C denaturing for 30 s 60 °C annealing for 30 s, and 72 °C extension for 1 min, followed by a final extension at 72 °C for 10 min. Quality and quantity of the purified PCR product was confirmed by 1·2% agarose gel electrophoresis. Antisense and sense RNA probes were labeled using the T7/T3 digoxigenin RNA labeling kit (Roche Diagnostics, Indianapolis, IN, USA), according to the manufacturer's instructions. Sections were pre-treated and hybridized with a Ventana RiboMap kit (Ventana Medical Systems) on the automated Ventana HX system Discovery. Detection of hybrids was performed with a digoxigenin nucleic acid detection kit (Boehringer Mannheim, Germany) following the manufacturer's instructions. Sections were then dehydrated through an ethanol series (80, 90 and 100% ethanol, for 1 min each) and washed for 1 min in xylene before mounting in malinol mounting medium (Muto Pure Chemicals Ltd, Tokyo, Japan).

Results and discussion

The expression profile of genes in pancreatic islets of experimental animals will greatly complement human studies of functional genomics of the tissue and the genetics of its disease states. This study both establishes a first molecular inventory of rat pancreatic islets and reveals a number of novel genes, the expression of which has not previously been described. These should provide important insights into the entire transcriptome of endocrine pancreas and be an immensely valuable aid to the improvement of genomic annotation.

Large-scale collection of ESTs from rat pancreatic islet and RINm5F cells

A total of 40 710 clones randomly selected from the two cDNA libraries were partially sequenced from the 3′-end. Sequences containing less than 1% ambiguous bases longer than 200 bp were subjected to BLASTN database search. Contaminated genomic sequences, e.g. repetitive sequences, (1967 clones) and mitochondrial DNAs (4633 clones), were removed from the pool of sequences, resulting in a collection of 34 110 ESTs comprising 22 310 known and 11 800 unknown sequences. Our previous study of 1000 ESTs from human pancreatic islets, the purity of which was

Table 1 Redundancy of 3′-ESTs from rat pancreatic islet and RINm5F cells

	Islet	RINm5F	Islet & RINm5F
Redundancy			
>1000	1	0	1
101–1000	1	3	7
51–100	3	6	15
31–50	6	12	28
21–30	21	31	69
11–20	87	138	253
6–10	228	299	533
3–5	1039	1113	1878
2	1556	1593	2075
1	3088	3065	5547

The 40 710 ESTs were generated by sequencing cDNA clones from the 3′-ends. After clustering analysis, a total of 6030 and 6260 independent groups were obtained from pancreatic islet and RINm5F cells, respectively.

estimated to be ~90% by microscopic examination and protein analysis, identified 13, 12, 6, and 9 ESTs for major exocrine genes for α-amylase, elastase, pancreatic lipase, and trypsinogen respectively (Takeda *et al.* 1993). Since only 2, 8, 2, and 5 clones for these exocrine genes respectively, were found in the ~20 000 EST sequences of this study, the possibility of contamination from exocrine cells appears to be quite negligible (~0·1%). This estimation of purity is consistent with that of the protein analysis described above. Because large-scale sequencing based on random isolation of clones generates high redundancy, clustering analysis was performed to assemble the sequences into non-redundant sequence groups. A total of 6030 and 6260 independent groups were obtained from pancreatic islet and RINm5F cells respectively, and the pattern of redundancy was similar between the two sources (Table 1 & 2). Together, 10 406 non-redundant sequences comprising 4859 clusters of sequences and 5547 singletons were obtained representing 4078 known genes and 6328 unknown genes. Since only 1896 distinct genes (18%) were found to overlap in the normal islets and RINm5F cells, this large-scale sequencing using two distinct cDNA sources was quite effective in identifying a larger number of non-redundant sequences. Studies of the number of different mRNA sequences in a cell suggest that a typical higher eukaryotic cell synthesizes 10 000 to 20 000 different proteins (Alberts *et al.* 1994), so this approach covered at least 50% of the possible protein-coding genes. As in similar large-scale cDNA sequencing studies carried out in other tissues, about 50% of the clones obtained are derived from genes not functionally annotated. These unknown clones were re-sequenced from the 5′-end, and the 5512 clones sequenced successfully were also subjected to database search. As a result, 1404 sequences representing 502 distinct transcripts showed perfect identity to known genes, so the 3′-end sequences of these clones clearly are not contained in the cDNA sequences deposited in the nucleotide databases. These ESTs were then assigned to the known group. All representative EST sequences obtained from each cluster were deposited in the public database to be freely available to all researchers (DDBJ accession No. BP464981–BP504629).

Table 2 Summary of non-redundant 3′-ESTs and/or 5′-ESTs

	Known	Unknown	Total
Islet only	1165	2981	4146
RINm5F only	1569	2807	4376
Islet & RINm5F	1344	540	1884
Total	4078	6328	10 406

The 6030 and 6260 non-redundant sequences from pancreatic islet and RINm5F cells, respectively, were subjected to BLASTN database search at the National Center for Biotechnology Information.

Characterization of known genes in pancreatic islet and RINm5F cells

The ESTs showing identity or high similarity to known genes were classified into seven major categories on the basis of putative general functions of the protein encoded, as described previously (categories: cell division, cell signaling/communication, cell structure/motility, cell/organism defense, gene/protein expression, metabolism, and unclassified). In total, 3951 out of 4078 known genes were represented in the classified data set (online supplement). The largest category of genes was gene/protein expression (26·4%). Successively smaller categories were cell signaling and communication (19·0%), metabolism (16·8%), cell structure/motility (7·7%), cell/organism defense (7·5%), and cell division (5·6%). ESTs lacking sufficient information to be classified constituted the remainder, unclassified (16·9%). To further analyze the molecular complexity, each major category was subdivided according to the putative specific functions of the proteins (Table 3, also see online supplement). For example, the largest category, gene/protein expression, was subdivided into eight subgroups. Of these, transcription factor constituted the largest number of non-redundant genes (416 genes by 1209 ESTs). The transcription factors include PDX-1, BETA2/NeuroD, HNF-4α, Nkx-2·2, Nkx-6·1, and Isl-1 etc, all of which are important for pancreatic development and islet-specific functions, and the first three of which are the causal genes for monogenic forms of diabetes, MODY4, MODY6, and MODY1 (Fajans *et al.* 2001). The other genes for transcription factors also are plausible candidates for diabetogenes or genes responsible for β-cell specific functions.

In this study, 60·8% of the non-redundant ESTs did not match any of the known genes in the nucleotide

Table 3 Functional categories of proteins encoded by non-redundant ESTs

Functional category	Subcategory	Islet only	RINm5F only	Islet & RINm5F	Total
Cell division	General	18	13	21	52
	DNA synthesis	1	18	5	24
	Apoptosis	13	15	18	46
	Cell cycle	13	36	19	68
	Chromosomal structure	8	15	9	32
	Subtotal	53	97	72	222
Cell signaling	Cell adhesion	20	18	12	50
	Channel	15	21	19	55
	Effectors	23	24	24	71
	Hormone	31	28	25	84
	Intracellular transducers	50	53	62	165
	Metabolism	2	1	3	6
	Protein modification	56	43	55	154
	Receptor	60	57	51	168
	Subtotal	257	245	251	753
Cell structure	General	9	17	16	42
	Contractile protein	9	8	7	24
	Cytoskeletal	23	26	39	88
	Extracellular matrix	24	10	8	42
	Microtubule-associated	17	23	17	57
	Vesicular transport	15	19	19	53
	Subtotal	97	103	106	306
Cell defense	Homeostasis (general)	21	16	20	57
	DNA repair	11	17	11	39
	Carrier protein	22	24	22	68
	Stress response	9	12	14	35
	Immunology	47	23	27	97
	Subtotal	110	92	94	296
Gene expression	RNA polymerase	6	6	12	24
	RNA processing	23	70	62	155
	Transcription factor	136	173	107	416
	Targeting	37	49	65	151
	Protein turnover	34	28	47	109
	Ribosomal proteins	17	20	87	124
	tRNA synthesis	1	9	8	18
	Translation factor	9	12	29	50
	Subtotal	263	367	417	1047
Metabolism	General	4	4	5	13
	Amino acid	12	15	18	45
	Cofactors	1	3	1	5
	Energy	31	37	56	124
	Lipid	39	51	49	139
	Nucleotide	18	30	13	61
	Protein modification	7	25	20	52
	Sugar	19	81	30	130
	Transport	25	36	31	92
	Subtotal	156	282	223	661
Unclassified		197	300	169	666
Total No. of unique genes		1133	1486	1332	3951

The ESTs showing identity or high similarity to known genes were classified into seven major categories on the basis of putative general functions of the protein encoded and each major category was further subdivided according to the putative specific functions of the proteins.

database. To identify novel rat genes encoding proteins structurally related to the known proteins, we performed BLASTX similarity search in the peptide databases using 5512 distinct ESTs. Of these, 127 represent rat homologs of genes identified in other species or new members of structurally related families in rat, the cut off for significant similarity being P value of 10^7 and similarity of 50% (Table 4). Functional analyses of the proteins encoded by these ESTs should clarify their novel roles in pancreatic islets.

Table 4 Rat homologs of known genes and new members of gene families

Clone ID	Gene	Species	%SIM	*P* value	Islet	RINm5F
RBC00545	breast carcinoma amplified sequence 3 homolog	Homo sapiens	0·72	2E-21	1	1
RBC00858	retrovirus-related POL polyprotein	Mus musculus	0·74	4E-42	1	0
RBC01084	eukaryotic translation initiation factor 4 gamma 3	Homo sapiens	0·66	5E-44	1	0
RBC01744	protein C20orf149 homolog	Rattus norvegicus	0·74	1E-51	1	0
RBC03066	DNA transformation protein comF	Pseudomonas stutzeri	0·74	3E-63	1	0
RBC03516	hypothetical protein in acoE 3′region	Rhodobacter sphaeroides	0·85	7E-48	1	0
RBC04593	speckle-type POZ protein-like 1; POZ 56 protein	Rattus norvegicus	0·72	3E-32	1	0
RBC05170	hypothetical protein 3	Rattus norvegicus	0·78	7E-35	1	1
RBC05610	14 kDa phosphohistidine phosphatase	Rattus norvegicus	0·64	9E-28	3	0
RBC06005	transposase for insertion sequence element IS904	Pseudomonas putida	0·83	2E-49	1	0
RBC06162	aurora-A kinase interacting protein	Rattus norvegicus	0·79	2E-45	2	0
RBC06680	epsin 1	Rattus norvegicus	0·75	8E-38	1	0
RBC07890	probable Pol polyprotein	Rattus norvegicus	0·71	5E-43	1	0
RBC08283	G protein-coupled receptor 150	Mus musculus	0·81	4E-89	1	0
RBC08942	Egl nine homolog 2	Mus musculus	0·75	2E-07	2	0
RBC09487	hypothetical protein PP2447 homolog	Mus musculus	0·75	4E-19	1	1
RBC09605	putative NF-kappa-B activating protein	Rattus norvegicus	0·64	1E-43	1	0
RBC10109	polyposis locus protein 1	Rattus norvegicus	0·79	2E-34	1	0
RBC11646	immunoglobulin light chain variable region	Mus musculus	0·58	2E-27	1	0
RBC12591	chromosome 10 open reading frame 45	Homo sapiens	0·62	4E-41	2	0
RBC12830	pORF2	Mus musculus	0·64	1E-34	1	0
RBC12840	methyl-accepting chemotaxis protein	Pseudomonas syringae	0·64	9E-34	1	0
RBC13044	ribonuclease P protein subunit p29	Homo sapiens	0·64	2E-19	2	0
RBC13272	hypothetical protein KIAA0174	Rattus norvegicus	0·76	5E-13	3	0
RBC13530	histone deacetylase 4 (HD4)	Rattus norvegicus	0·73	2E-18	1	0
RBC14441	lactoylglutathione lyase	Rattus norvegicus	0·71	7E-49	2	0
RBC14577	lymphocyte antigen Ly-6D precursor	Rattus norvegicus	0·79	4E-38	2	0
RBC15029	protein C21orf5	Homo sapiens	0·85	7E-33	1	0
RBC15417	POL polyprotein	Trichosurus vulpecula	0·71	9E-25	2	0
RBC15535	amino-acid ABC transporter ATP-binding protein	Thermobifida fusca	0·83	3E-80	1	0
RBC15661	zinedin	Rattus norvegicus	0·76	1E-22	1	1
RBC15731	collagen alpha 1(X) chain	Mus musculus	0·66	3E-64	4	0
RBC17834	probable transcriptional regulatory protein ygiX	synthetic construct	0·72	2E-53	1	0
RBC18842	nectin 4	Mus musculus	0·75	8E-73	1	0
RBC19899	probable 3-hydroxybutyryl-CoA dehydrogenase	Brucella melitensis	0·81	3E-99	3	0
RBC20402	hypothetical protein CGI-143	Rattus norvegicus	0·64	4E-28	1	0
RIN00207	S-adenosylmethionine-dependent methyltransferase	Mycoplasma mobile	0·68	3E-54	0	2
RIN00228	heat inducible transcriptional repressor protein	Mycoplasma mobile	0·59	3E-32	0	1
RIN00249	cell cycle control protein cwf15	Mus musculus	0·56	4E-12	0	1
RIN00488	hypothetical protein C5D6·06c	Mus musculus	0·86	3E-85	0	1
RIN00804	zinc finger protein 35 (Zfp-35)	Rattus norvegicus	0·62	5E-22	0	1
RIN01128	tyrosine-protein kinase FLK	Rattus norvegicus	0·77	1E-42	0	1
RIN01130	trigger factor (TF)	Mycoplasma pulmonis	0·65	7E-11	0	1
RIN01171	WD-repeat protein CGI-48	Rattus norvegicus	0·6	9E-28	0	1
RIN01386	probable cation-transporting ATPase 1	Rattus norvegicus	0·79	3E-82	0	1
RIN01389	L-lactate dehydrogenase	Mesoplasma florum L1	0·58	2E-20	0	1
RIN01548	THAP domain protein 11	Rattus norvegicus	0·51	9E-22	0	1
RIN01784	hypothetical protein KIAA0233	Rattus norvegicus	0·86	5E-23	0	1
RIN01794	cell division protein ftsH homolog	Mycoplasma pulmonis	0·64	4E-32	0	1
RIN01840	hypothetical protein MG061	Mycoplasma pulmonis	0·55	1E-34	0	4
RIN01932	condensin complex subunit 2 (p105)	Mus musculus	0·88	3E-14	0	1
RIN02139	protein FAM3A precursor	Mus musculus	0·87	5E-36	0	1
RIN02177	ABC transporter ATP-binding protein	Mycoplasma gallisepticum	0·74	7E-60	0	3
RIN02445	translation initiation factor IF3	Mycoplasma fermentans	0·62	8E-12	0	2
RIN02483	splicing factor, arginine/serine-rich 2	Mus musculus	0·66	2E-59	0	2

Table 4 *Continued*

Clone ID	Gene	Species	%SIM	*P* value	Islet	RINm5F
RIN02532	ABC transporter permease protein MG188 homolog	Mycoplasma gallisepticum	0·65	7E-11	0	2
RIN02603	protein c20orf172 homolog	Mus musculus	0·89	2E-44	0	1
RIN03035	probable nicotinate-nucleotide adenylyltransferase	Mycoplasma mobile	0·65	3E-26	0	1
RIN03110	phospholipase A2 inhibitor gamma subunit B	Rattus norvegicus	0·84	2E-07	0	3
RIN03339	hypothetical lipoprotein MPN288	Mycoplasma gallisepticum	0·56	4E-10	0	2
RIN03427	putative Pol polyprotein	Rattus norvegicus	0·77	2E-19	0	2
RIN04062	hypothetical protein MG148	Ureaplasma parvum	0·69	3E-22	0	1
RIN04617	probable cation-transporting P-type ATPase	Ureaplasma parvum	0·55	1E-31	0	1
RIN05010	testis-specific protein PBS13	Rattus norvegicus	0·78	5E-88	0	1
RIN05122	hypothetical protein KIAA0036	Rattus norvegicus	0·86	6E-76	0	1
RIN05229	E3 ubiquitin-protein ligase Nedd-4	Cricetulus griseus	0·86	4E-48	0	1
RIN05576	TLM protein	Mus musculus	0·64	1E-26	2	1
RIN05605	RRS1 ribosome biogenesis regulator homolog	Mus musculus	0·67	3E-35	0	1
RIN05994	eukaryotic translation initiation factor 4E transporter	Rattus norvegicus	0·7	3E-13	0	1
RIN06079	brain protein 14	Mus musculus	0·61	8E-12	0	1
RIN06155	slingshot 3	Rattus norvegicus	0·79	6E-47	0	1
RIN06262	mitochondrial respiratory chain complexes assembly protein	Mycoplasma pulmonis	0·73	3E-37	0	1
RIN06288	SCO2 protein homolog, mitochondrial precursor	Mus musculus	0·76	4E-45	0	3
RIN06497	methionine aminopeptidase	Mycoplasma pulmonis	0·7	1E-37	0	1
RIN06650	retrovirus-related protease	Homo sapiens	0·54	2E-12	0	1
RIN07055	TGF-beta induced apoptosis protein 2	Rattus norvegicus	0·62	2E-51	0	1
RIN07707	methyl-CpG binding domain protein 6	Rattus norvegicus	0·61	1E-24	0	1
RIN07765	splicing factor 3 subunit 1	Mus musculus	0·55	2E-35	0	1
RIN07973	NADH-ubiquinone oxidoreductase chain 4L	Rattus norvegicus	0·72	7E-23	0	1
RIN08083	signal recognition particle protein	Mycoplasma mobile	0·72	1E-21	0	1
RIN08132	forkhead box protein K1	Homo sapiens	0·66	1E-26	0	1
RIN08211	meningioma-expressed antigen 6/11	Rattus norvegicus	0·71	2E-27	2	1
RIN08412	hypothetical 45·0 kDa protein in NOT1-MATAL2 region	Rattus norvegicus	0·85	3E-79	0	1
RIN09109	TED protein	Mus musculus	0·87	5E-85	0	3
RIN09205	SRR1-like protein	Homo sapiens	0·83	2E-42	0	4
RIN09341	30S ribosomal protein S6	Mycoplasma pulmonis	0·55	2E-17	0	7
RIN09690	valyl-tRNA synthetase	Mycoplasma pulmonis	0·87	5E-40	0	5
RIN10182	protein transport protein SEC61 gamma subunit	Rattus norvegicus	0·7	5E-20	0	4
RIN10530	hypothetical protein KIAA0893	Mus musculus	0·81	9E-93	1	3
RIN10621	splicing factor, arginine/serine-rich 8	Mus musculus	0·51	5E-15	0	1
RIN10814	G protein-coupled receptor family C group 5 member C	Rattus norvegicus	0·73	8E-43	0	2
RIN11008	tRNA-splicing endonuclease subunit SEN54	Rattus norvegicus	0·86	6E-98	0	3
RIN11027	protein yhgF	Homo sapiens	0·89	9E-93	1	1
RIN11030	p53-associated parkin-like cytoplasmic protein	Rattus norvegicus	0·84	9E-44	0	1
RIN11408	protein disulfide isomerase precursor	Rattus norvegicus	0·74	7E-35	0	1
RIN11741	polyhomeotic-like protein 1	Rattus norvegicus	0·82	3E-54	0	2
RIN11820	translation initiation factor IF-1	Leptospira interrogans	0·74	9E-12	0	1
RIN11856	fructose-bisphosphate aldolase	Mycoplasma mobile	0·8	6E-52	0	1
RIN11899	nuclear protein Hcc-1	Mus musculus	0·72	2E-67	0	2
RIN12027	60S ribosomal protein L22	Mus musculus	0·81	2E-39	0	1
RIN12420	putative ATP-dependent RNA helicase T26G10	Rattus norvegicus	0·8	4E-83	0	2
RIN12461	synaptic vesicle membrane protein VAT-1 homolog	Rattus norvegicus	0·73	8E-65	0	3
RIN12488	glutaminyl-peptide cyclotransferase	Rattus norvegicus	0·78	7E-90	0	2
RIN12608	probable RNA-dependent helicase p72	Rattus norvegicus	0·64	8E-62	0	2
RIN12715	oligoendopeptidase F homolog	Mycoplasma pulmonis	0·68	8E-58	0	1
RIN12733	PSL10 protein	Mus musculus	0·77	7E-80	0	1
RIN12794	hypothetical protein ORF-1137	Rattus norvegicus	0·72	2E-32	0	1
RIN12806	protein HI1455	Mycoplasma pulmonis	0·74	7E-50	0	3

Table 4 *Continued*

Clone ID	Gene	Species	%SIM	*P* value	Islet	RINm5F
RIN12897	dimethyladenosine transferase	Mycoplasma gallisepticum	0·67	5E-36	0	2
RIN12933	zinc finger protein 23	Mus musculus	0·79	4E-63	1	2
RIN13293	serine hydroxymethyltransferase	Rattus norvegicus	0·65	4E-93	0	1
RIN13306	peripherin	Mus musculus	0·7	1E-42	0	1
RIN13714	seryl-tRNA synthetase	Mycoplasma pulmonis	0·82	3E-62	0	1
RIN13766	UBA/UBX 33·3 kDa protein	Rattus norvegicus	0·65	1E-54	0	1
RIN13865	splicing factor 1	Mus musculus	0·73	6E-23	0	2
RIN14137	condensin subunit 1	Mus musculus	0·62	2E-33	0	1
RIN14394	Csr1	Cricetulus griseus	0·82	8E-12	0	1
RIN14494	hypothetical protein KIAA0117	Rattus norvegicus	0·77	7E-81	0	1
RIN14758	general transcription factor 3C polypeptide 5	Rattus norvegicus	0·73	1E-48	1	1
RIN15622	Crumbs protein homolog 1	Mus musculus	0·6	7E-08	0	1
RIN16005	hypothetical protein C19A8·09	Rattus norvegicus	0·89	4E-11	0	2
RIN16184	single-strand binding protein 1	Rattus norvegicus	0·67	6E-13	1	1
RIN16330	zinc finger protein 510	Mus musculus	0·83	2E-19	0	2
RIN16398	60S ribosomal protein L7	Rattus norvegicus	0·83	1E-32	1	3
RIN16446	single-strand binding protein	Rattus norvegicus	0·87	3E-13	0	1
RIN16588	putative adenosylhomocysteinase 2	Homo sapiens	0·89	2E-07	0	2
RIN16592	thymidine phosphorylase	Mycoplasma pulmonis	0·66	3E-20	0	1

The unknown clones after BLASTN search in the nucleotide databases were re-sequenced from the 5′-end and then subjected to BLASTX search in the peptide databases. The cut off used for significant similarity was *P* value of 10^{-7} and similarity of 50%. The clone IDs RBC and RIN show ESTs from pancreatic islet and RINm5F cells, respectively.

To identify genes that had not been determined to be expressed, the sequences showing no significant match to any of the annotated genes were further compared with dbEST entries from other tissues, revealing 896 expressed genes that have not appeared in the database. These 896 putatively novel genes were analyzed in the context of recently determined rat genomic sequences (Rat Genome Sequencing Project Consortium 2004). Of these genes, 779 were encoded by exon-intron organization in the corresponding genomic sequences. However, since the transcripts of many of these genes were barely detectable by *in situ* hybridization, they may be expressed at low levels, at least in adult islets. Because chemiluminescence-based *in situ* detection of mRNAs is not sufficiently sensitive, a large-scale RT-PCR analysis is presently in progress in our laboratory to elucidate the tissue distribution. The expression of 71 of the other 117 sequences was uncertain due to their ambiguous genomic structure. The remainder could be derived from possible pseudogenes or retroposons, as the corresponding sequences in the genome are flanked by AT-rich sequences that were recognizable by oligo-dT priming in the process of cDNA synthesis.

Characterization of differentially expressed genes

The immunoreactive insulin (IRI) content of rat normal β-cells has been reported to be ~8000 pmol/10^6 cells, while the RINm5F cell line used in this study has been estimated to contain a very low level of IRI (0·43 pmol/10^6 cells) and a much lower number of secretory granules (Kayo *et al.* 1997). Accordingly, the expression levels of the genes involved in insulin synthesis and secretion in RINm5F cells should markedly differ from those of the normal β-cells. In addition, since the RINm5F cells were derived from radiation-induced tumor cells and exhibited a decrement of well-differentiation (Gazdar *et al.* 1980, Philippe *et al.* 1987, Kayo *et al.* 1997), the expression levels of the genes involved in cell differentiation and tumorigenesis may also be altered. The relative frequencies of ESTs have been shown to reflect the average level of expression of the corresponding mRNAs in the tissues examined (Lee *et al.* 1995). As pancreatic islet cells are mostly β-cells, the expression profile of insulin-related genes in the two cDNA sources (of same size) can be compared to identify differentially expressed genes. The EST frequencies for most of the house-keeping genes were similar in the two cDNA libraries, suggesting that such comparison of EST frequencies is reasonable. Over 2-fold differences in frequency between the two libraries were found in 204 genes (higher EST>10 times). The direction of change in mRNA levels in these ESTs observed by comparison of the EST frequency and the TaqMan semi-quantitative analysis was quite parallel, while the magnitude of the change was not correlated. The representative results for some of the ESTs (>15 times) are shown in Fig. 1. Previously, similar comparative analysis using ~6000 ESTs from two different conditions of PC-12 cells was performed (Lee *et al.* 1995). The study found the ratio of EST frequencies between the two cDNA sources to be