Diagnostic Pathology

Research

Open Access

Theory of sampling and its application in tissue based diagnosis

Klaus Kayser*[1], Holger Schultz[2], Torsten Goldmann[2], Jürgen Görtler[3], Gian Kayser[4] and Ekkehard Vollmer[2]

Address: [1]UICC-TPCC, Institute of Pathology, Charite, Berlin, Germany, [2]Clin. & Exp. Pathology, Research Center Borstel, Borstel, Germany, [3]Deep Computing, IBM, Amsterdam, the Netherlands and [4]Institute of Pathology, University of Freiburg, Freiburg, Germany

Email: Klaus Kayser* - klaus.kayser@charite.de; Holger Schultz - hschultz@fz-borstel.de; Torsten Goldmann - tgoldmann@fz-borstel.de; Jürgen Görtler - goertler@de.ibm.com; Gian Kayser - gian.kayser@uniklinik-freiburg.de; Ekkehard Vollmer - evollmer@fz-borstel.de

* Corresponding author

Published: 16 February 2009

Diagnostic Pathology 2009, **4**:6 doi:10.1186/1746-1596-4-6

Received: 27 January 2009
Accepted: 16 February 2009

This article is available from: http://www.diagnosticpathology.org/content/4/1/6

Abstract

Background: A general theory of sampling and its application in tissue based diagnosis is presented. Sampling is defined as extraction of information from certain limited spaces and its transformation into a statement or measure that is valid for the entire (reference) space. The procedure should be reproducible in time and space, i.e. give the same results when applied under similar circumstances. Sampling includes two different aspects, the procedure of sample selection and the efficiency of its performance. The practical performance of sample selection focuses on search for localization of specific compartments within the basic space, and search for presence of specific compartments.

Methods: When a sampling procedure is applied in diagnostic processes two different procedures can be distinguished: I) the evaluation of a diagnostic significance of a certain object, which is the probability that the object can be grouped into a certain diagnosis, and II) the probability to detect these basic units. Sampling can be performed without or with external knowledge, such as size of searched objects, neighbourhood conditions, spatial distribution of objects, etc. If the sample size is much larger than the object size, the application of a translation invariant transformation results in Kriege's formula, which is widely used in search for ores. Usually, sampling is performed in a series of area (space) selections of identical size. The size can be defined in relation to the reference space or according to interspatial relationship. The first method is called random sampling, the second stratified sampling.

Results: Random sampling does not require knowledge about the reference space, and is used to estimate the number and size of objects. Estimated features include area (volume) fraction, numerical, boundary and surface densities. Stratified sampling requires the knowledge of objects (and their features) and evaluates spatial features in relation to the detected objects (for example grey value distribution around an object). It serves also for the definition of parameters of the probability function in so – called active segmentation.

Conclusion: The method is useful in standardization of images derived from immunohistochemically stained slides, and implemented in the EAMUS™ system http://www.diagnomX.de. It can also be applied for the search of "objects possessing an amplification function", i.e. a rare event with "steering function". A formula to calculate the efficiency and potential error rate of the described sampling procedures is given.

Introduction

Diagnostic surgical pathology or tissue – based diagnosis is confronted with remarkable changes in its environment and workflow. The technological progress has led to a broad application of molecular biological methods such as Fluorescent in Situ Hybridization (FISH), and other DNA – sequence amplification techniques [1,2]. Commercially available slide scanners digitize a complete glass slide within a few minutes, and permit the implementation of completely digitized images into routine diagnostics [3,4]. In other words, the workload of a pathologist increases steadily not only by increase of material, but, in addition, due to the mandatory introduction of new, still tissue – based diagnostic technologies. Thus, the question arises: How can the availability of and access to digitized histological slides (virtual slides) be used to release the diagnostic pathologist from time consuming work steps in order to make the pathologist's work more effective and disease related?

In the early days of telepathology, which can be considered to be the "mother of the digital pathologist's world", several authors reported on the diagnostic accuracy of viewing digitized slides in comparison to conventional microscopy [4-8]. The results were clear: the diagnostic accuracy viewing at a digitized (or virtual) slide is indistinguishable to that of conventional microscopy; however, the required time is essentially longer [9,10]. The non appropriate and more time consuming search for appropriate fields of view or the performed sampling procedure are obviously one reason of these constraints. To our knowledge, the theory of sampling in cytology and histopathology has not been described in detail, and is nearly unknown in the environment of diagnostic pathologists. In this article we want to explain the main theoretical aspects and the derivatives of sampling which are performed in routine tissue – based diagnostics. The derived formulas will allow interested pathologists or scientists to search for applications that can diminish the sampling time in virtual slides.

Basic aspects of sampling in digitized histological slides (virtual slides)

Surgical pathology is a medical discipline that "extracts" information from human tissue and classifies the information in distinct terms that are called diagnoses. The common performance is to screen an organ or a tissue section for those spaces or areas that contain the most significant information, and try to classify this information seen in the specific field of view. Thus, tissue – based diagnosis is based upon a procedure to search for small samples that allow to derive information that is valid for the whole (or even patient). In other words, an appropriate sampling procedure is a precondition to evaluate accurate and reproducible diagnoses [2,4,11-15]. Therefore, a detailed definition and accurate description of the sampling method is a necessity if we want to further evaluate the diagnostic algorithms. This statement induces the definition of sampling as follows: Sampling is a method to derive information from a limited (small) compartment of a large (even unlimited) system that is valid for the entire (basic) system. The system can be a space, a function or set of functions, a body, an organ, a slide, or a DNA sequence.

The definition includes the term information, which has again to be defined: Information is a property that is exchanged between a sender and a receiver. Information is a property that can be understood by, and allows the receiver to react in an adequate, i.e., predictive manner. This definition of sampling includes two different aspects, which depend upon each other:

1. the method of sampling, and

2. the aim of the sampling procedure, i.e., which information should be extracted.

Different aims can require different methods of sampling, or at least different parameters of the same algorithm. The inclusion of an "aim" or "goal" to be assessed introduces the calculation of efficiency, or a cost/benefit estimation.

The most frequently used sampling goals are

➢ search for localization of specific items within the basic space, with the knowledge or assumption, that the space under consideration contains such items, and

➢ search for presence of specific items (tumour cells, ores, lobster, etc.), where the exact localisation of these items is of minor interest (for example localization of tumour cells in a cytological smear).

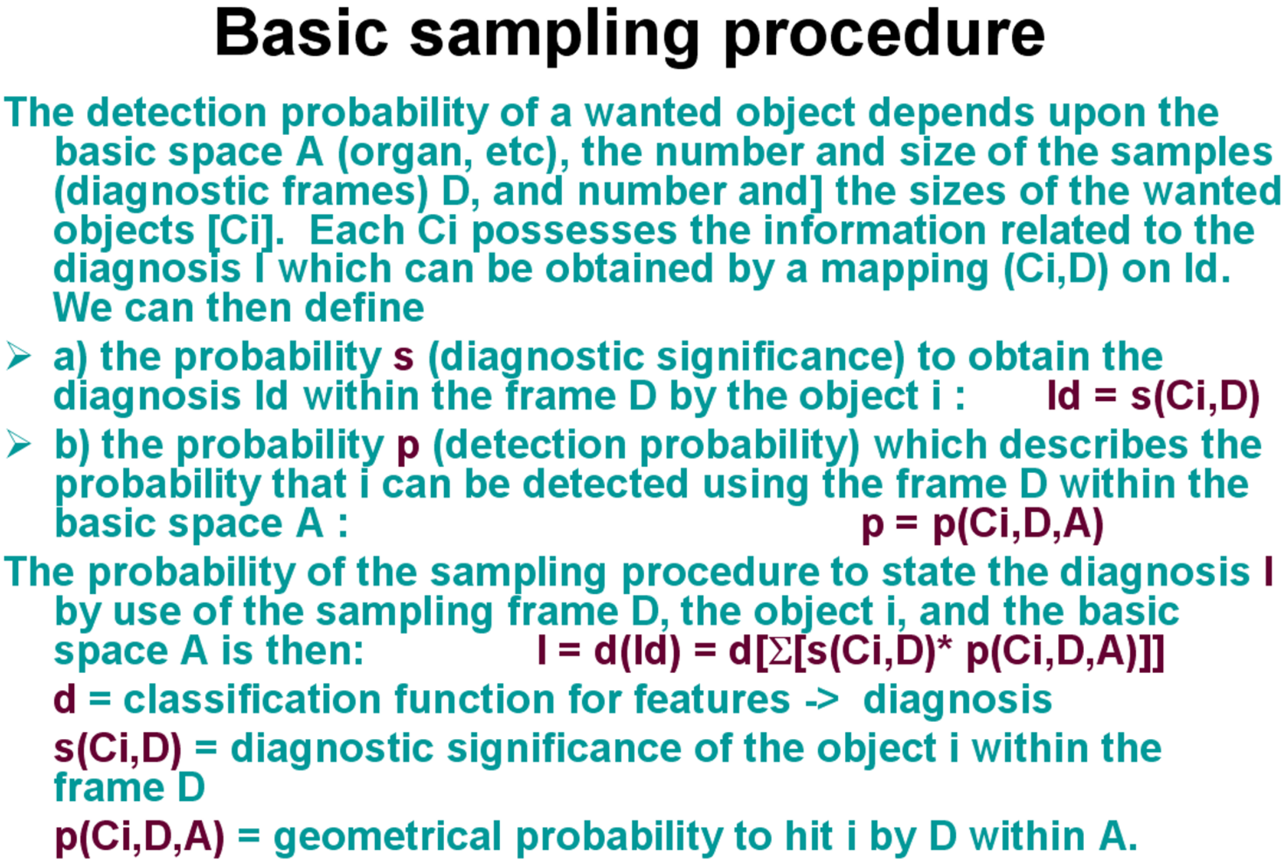
The prepositions to apply an adequate sampling procedure in tissue – based diagnosis include that number and size of the samples are limited. In addition, the detectable information has to be known. This information commonly depends upon additional (external) factors, and can be translated into diagnostic features that allow the detection and identification of a probe within the sampling space. These features can depend upon the size of the probes, their number, and their position within the collective, or even within the sampling space.

Let us assume that the final goal of our sampling method is the extraction of information from the entire space, and the classification of this information into a diagnosis. The diagnostic process can be separated into two different procedures:

I) the evaluation of a diagnostic significance of a certain object or "basic unit" which is the probability that the object can be grouped into a certain diagnosis, and

II) the probability to detect these basic units within the entire space.

The detection probability of a wanted object depends then upon the size of the basic space A (filed of view, organ, nucleus, etc), the number and size of the samples (diagnostic frames) D, and number and the sizes of the detected objects [Ci], as demonstrated in figure 1. Each detected object Ci possesses a certain probability to contribute to the diagnosis I which can be obtained by a mapping (Ci, D) on I_D. I_D is the probability to state the diagnosis I within the frame D by the object i, or ID = s(Ci, D). Basic examples are given in figure 2, figure 3, and in figure 4. In principle, the differentiation of the mapping of (Ci, D) on I_D into the two different procedures, namely a) the detection (geometric significance) and into the diagnostic contribution reflects to the applied segmentation algorithms. These distinguish between areas (pixels) that contributed to the object and those that do not. Measurements of objects can only be done if the object is completely covered by the sample frame. The spatial selection of the samples can either be performed randomly, or dependent upon the localization of already segmented objects (stratified sampling). Stratified sampling is based upon the general law of self organization, i.e., similar biological systems tend to be localized in a neighbourhood relation. In other words, to detect a cancer cell is more likely in the neighbourhood of an already identified cancer cell than elsewhere. The function of diagnostic significance is often of exponential nature, and in light microscopy related to three different image properties,



Basic sampling procedure

The detection probability of a wanted object depends upon the basic space A (organ, etc), the number and size of the samples (diagnostic frames) D, and number and] the sizes of the wanted objects [Ci]. Each Ci possesses the information related to the diagnosis I which can be obtained by a mapping (Ci,D) on Id. We can then define

- a) the probability s (diagnostic significance) to obtain the diagnosis Id within the frame D by the object i : Id = s(Ci,D)
- b) the probability p (detection probability) which describes the probability that i can be detected using the frame D within the basic space A : p = p(Ci,D,A)

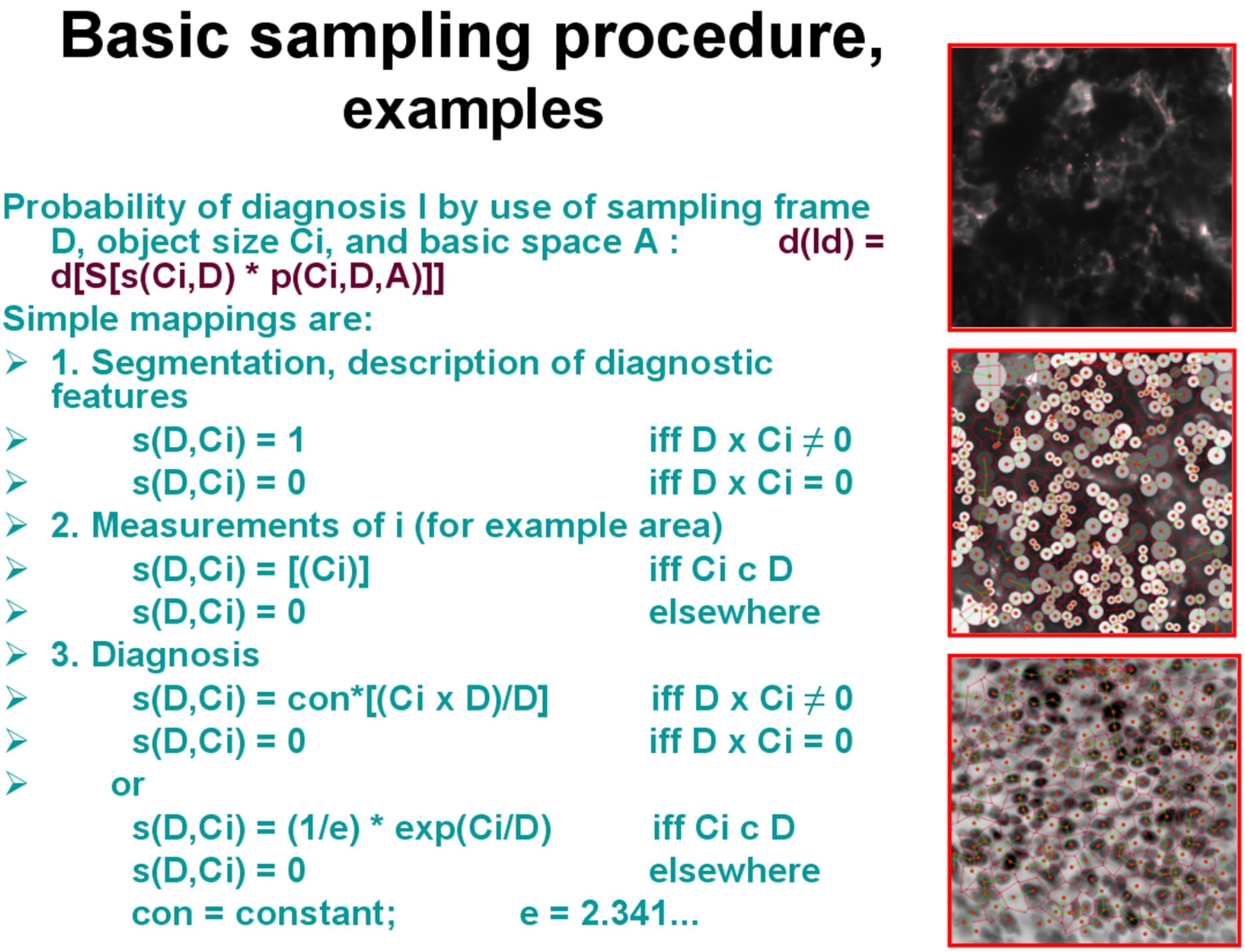
The probability of the sampling procedure to state the diagnosis I by use of the sampling frame D, the object i, and the basic space A is then: I = d(Id) = d[Σ[s(Ci,D)* p(Ci,D,A)]]

d = classification function for features -> diagnosis

s(Ci,D) = diagnostic significance of the object i within the frame D

p(Ci,D,A) = geometrical probability to hit i by D within A.

Figure 1
Survey of sampling algorithm.



Basic sampling procedure, examples

Probability of diagnosis I by use of sampling frame D, object size Ci, and basic space A : d(Id) = d[S[s(Ci,D) * p(Ci,D,A)]]

Simple mappings are:

- 1. Segmentation, description of diagnostic features
- s(D,Ci) = 1 iff D x Ci ≠ 0
- s(D,Ci) = 0 iff D x Ci = 0
- 2. Measurements of i (for example area)
- s(D,Ci) = [(Ci)] iff Ci c D
- s(D,Ci) = 0 elsewhere
- 3. Diagnosis
- s(D,Ci) = con*[(Ci x D)/D] iff D x Ci ≠ 0
- s(D,Ci) = 0 iff D x Ci = 0
- or

s(D,Ci) = (1/e) * exp(Ci/D) iff Ci c D

s(D,Ci) = 0 elsewhere

con = constant; e = 2.341...

Figure 2
Examples of different sampling procedures.

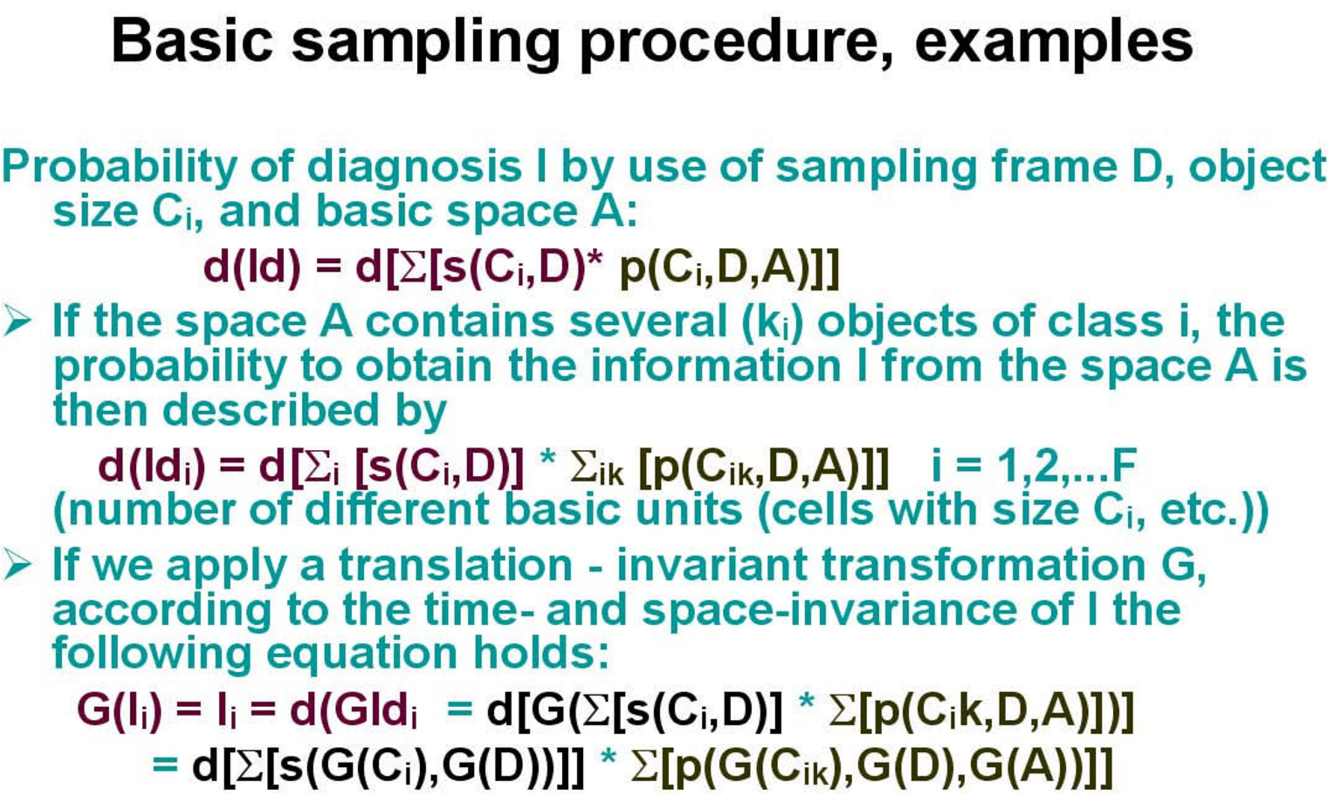
namely the texture, the object, and the object – associated structure [12,16-19].

Sampling aims, applications and examples

Sampling is basically an information detection and transformation procedure, and thus undertaken to reach a certain final aim, for example to state a diagnosis, or to identify the presence or absence of certain objects. A time and space invariant translation of the sampling procedure can be assumed as long as we want to obtain reproducible results (figure 3). Such a translation permits a separation of the object detection likelihood from the diagnostic significance of the segmented objects, and allows us to compute both properties separately. Assuming a digitized image, each point (pixel x, y) within the basic image is either a presentation of an object or not. All object features can be reduced to a function that represents the object pixels in relation to sample and basic image size (figure 4). The introduction of an exponential diagnosis function then gives us the well known formula of Krige, which is commonly used to detect ores, oil fields, or underground water reservations [20]. Furthermore, the application of specific mappings (dilation, erosion) permits us to "increase the magnification of an object within its sampling frame, or to define the center of gravity in an object in order to compute the image structure. One will obtain a so – called order of structures if these procedures are repetitively applied to light microscopy [21].

Random and stratified sampling procedures

The prerequisite of any (random, stratified) sampling is at least a binary image, i.e. a foreground defining the basic units and a background have to exist. Any sampling procedure can be performed either as random or stratified sampling figure 5 and figure 6. In addition, an active, passive, and a functional sampling can be distinguished.



Basic sampling procedure, examples

Probability of diagnosis I by use of sampling frame D, object size C_i, and basic space A:

$$d(Id) = d[\Sigma[s(C_i,D)^* \; p(C_i,D,A)]]$$

➢ If the space A contains several (k_i) objects of class i, the probability to obtain the information I from the space A is then described by

$$d(Id_i) = d[\Sigma_i \, [s(C_i,D)] \, * \, \Sigma_{ik} \, [p(C_{ik},D,A)]] \quad i = 1,2,...F$$

(number of different basic units (cells with size C_i, etc.))

➢ If we apply a translation - invariant transformation G, according to the time- and space-invariance of I the following equation holds:

$$G(I_i) = I_i = d(GId_i \;\; = d[G(\Sigma[s(C_i,D)] \, * \, \Sigma[p(C_ik,D,A)])]$$
$$= d[\Sigma[s(G(C_i),G(D))]] \, * \, \Sigma[p(G(C_{ik}),G(D),G(A))]]$$

Figure 3
Examples of probability calculations in various sampling procedures.

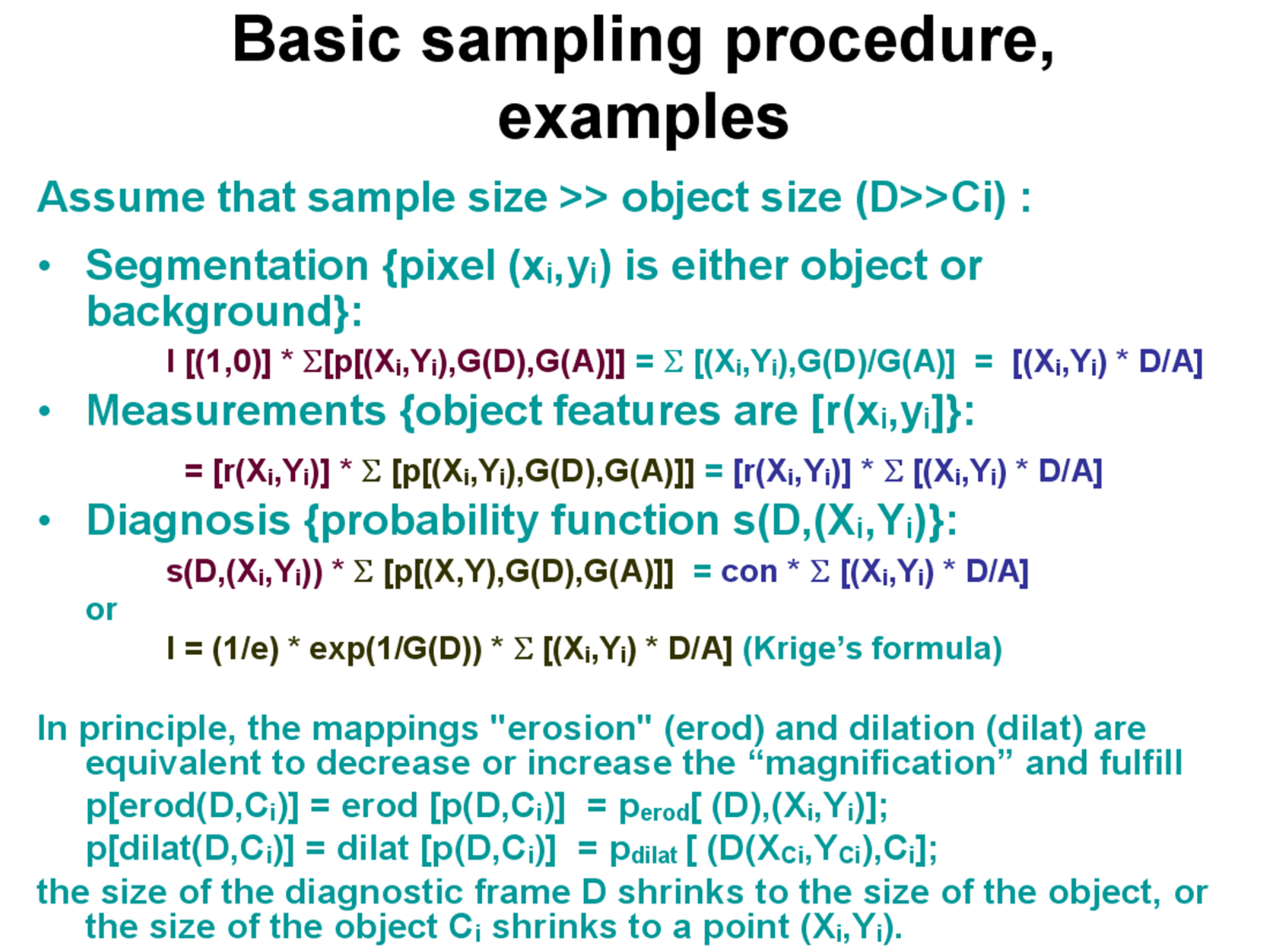
Random sampling is the selection of biological meaningful units (nuclei, cells, proteins, etc.) at random. It is used to measure

➢ the frequency of the analyzed units in relation to each other or to the basic space (structure);

➢ certain features of the biological units in relation to the basic space (to further identify and classify the objects).

No information about the basic (reference) space is needed. The detection of biological meaningful units is then equivalent to the segmentation of the image and analysis of randomly chosen segmented elements. This procedure forms the basis of numerous investigations since the 1950s. It is commonly called stereology [22-26]. In principle, a grid consisting of regular lines (points) with identical length (and distance in between) is overlayed to the image, and the number of hits (intersections) is counted. From the number of intersections the volume – adjusted frequency, size, surface can be derived, independent from the orientation and shape of the elements. In a binary image the pixels (binary x, y points) can be used as a grid. Random chosen are the cutting angle (plane/volume), and the start point of the grid (pixel). Stratified are the selection of the grid (all pixels) and the count of intersections. Thus, any random sampling is provided by the start of the procedure, for example by random selection of the upper right position (x, y) coordinates of the sample space. From the relation *x/A* (number of hits *x/reference area A*) two-dimensional (and also three – dimensional) parameters can be derived. These include the area density (*Aa*), the volume density (*Vv*), the boundary density (*Ba*), the numerical density (*Na*), and the surface density (*Sv*). It should be noted, that this quite easily applied procedure permits the estimation of significant three – dimensional object features without any sophisticated three dimensional reconstruction [23,27,18,29].

Stratified sampling, in contrast, is provided by a specific selection of intersections (objects). Its objective is the detection of specific objects (of known features), and the measurement of features of known specific objects, or the estimation of objective-associated reference volumes, for example the density of proliferating cells related to distance from the nearest vessel [3,9,18,21]. Using again a grid as a measurement tool, the cutting angle (plane/volume) and the start point of the grid (pixel) are also randomly chosen. Stratified are the selection of the grid (all



Basic sampling procedure, examples

Assume that sample size >> object size (D>>Ci) :

- Segmentation {pixel (x_i,y_i) is either object or background}:

$$I\ [(1,0)] * \Sigma[p[(X_i,Y_i),G(D),G(A)]] = \Sigma\ [(X_i,Y_i),G(D)/G(A)] = [(X_i,Y_i) * D/A]$$

- Measurements {object features are [r(x_i,y_i]}:

$$= [r(X_i,Y_i)] * \Sigma\ [p[(X_i,Y_i),G(D),G(A)]] = [r(X_i,Y_i)] * \Sigma\ [(X_i,Y_i) * D/A]$$

- Diagnosis {probability function s(D,(X_i,Y_i)}:

$$s(D,(X_i,Y_i)) * \Sigma\ [p[(X,Y),G(D),G(A)]] = con * \Sigma\ [(X_i,Y_i) * D/A]$$

or

$I = (1/e) * \exp(1/G(D)) * \Sigma\ [(X_i,Y_i) * D/A]$ (Krige's formula)

In principle, the mappings "erosion" (erod) and dilation (dilat) are equivalent to decrease or increase the "magnification" and fulfill

$p[erod(D,C_i)] = erod\ [p(D,C_i)] = p_{erod}[\ (D),(X_i,Y_i)];$

$p[dilat(D,C_i)] = dilat\ [p(D,C_i)] = p_{dilat}\ [\ (D(X_{Ci},Y_{Ci}),C_i];$

the size of the diagnostic frame D shrinks to the size of the object, or the size of the object C_i shrinks to a point (X_i,Y_i).

Figure 4
Examples of sampling procedures in segmentation and diagnosis classification.

pixels), and the count of specific intersections (for example large cells) only. Thus: stratified sampling is provided by specific a selection of intersections (objects). A classic example is its application in cytology, i.e. to find the diagnosis-relevant cell (tumor cell) within a large number of "normal" cells. One could try to analyze

1. only those areas which contain features of (any) cell (gray value selection at low magnification)

2. within these areas only those cells which seem to be abnormal (gray value, size, moderate magnification)

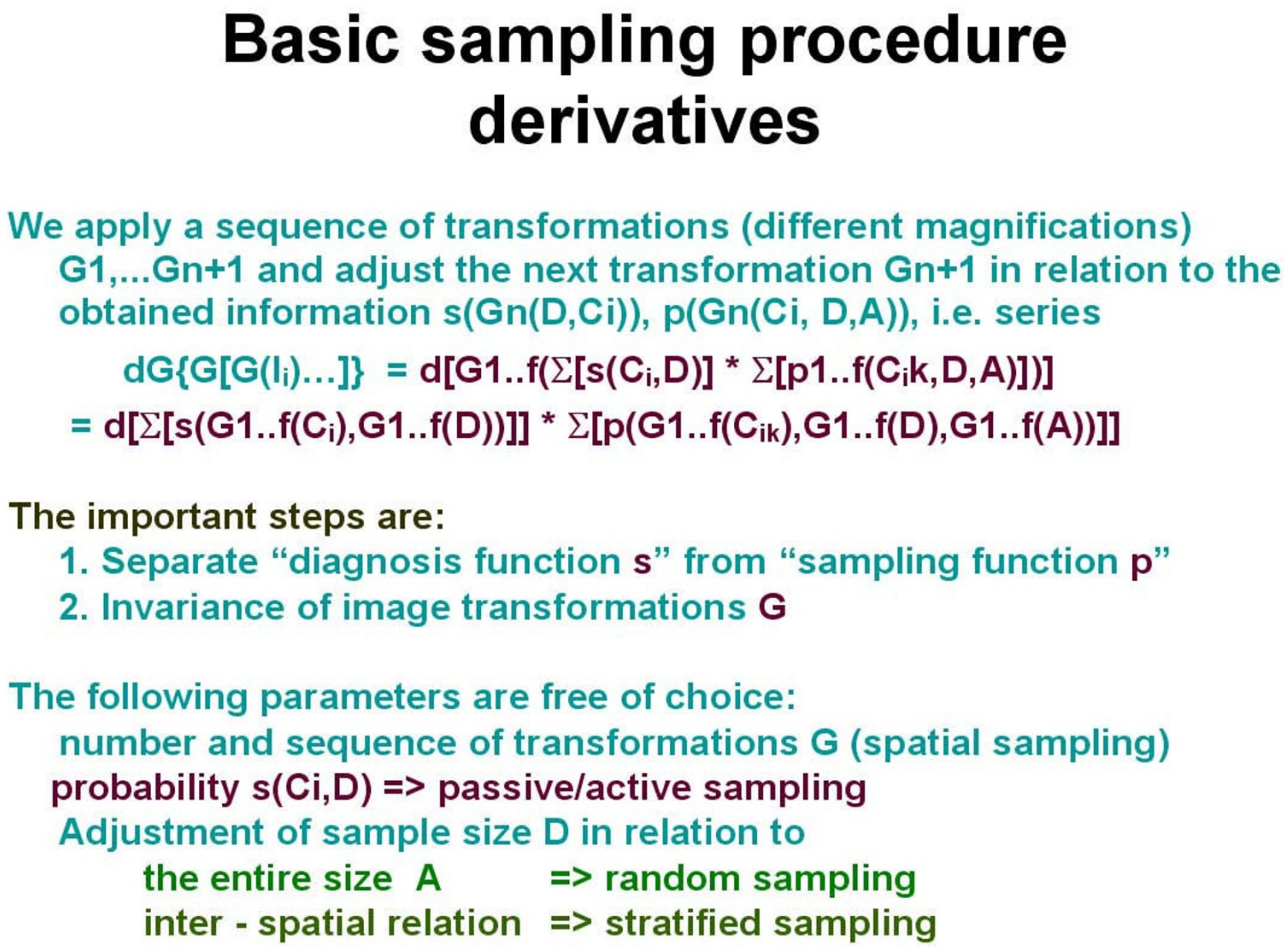
3. within these cells those with abnormal nuclear size (DNA content), high magnification.

4. terminate the procedure once the diagnosis – significant information has been obtained

All other items are disregarded or neglected. The implementation of such an algorithm can speed up the time required "to screen a slide" significantly [5,12,30].

Stratified sampling requires some external knowledge in order to detect the biological meaningful events such as cancer cells. The image features of a cancer cell have to be known if one would like to detect this event by stratified sampling. The alternative algorithm would be to "sample" all cells, and start, if possible, a statistical analysis. This would then try to evaluate the rare events (supposing that cancer cells are rare to normal cells). Again, some external knowledge would be necessary. Obviously, this is related to the diagnosis function s(Ci, D).

Stratified sampling requires an accurate segmentation of objects with known features. Independent upon the



Basic sampling procedure derivatives

We apply a sequence of transformations (different magnifications) G1,...Gn+1 and adjust the next transformation Gn+1 in relation to the obtained information s(Gn(D,Ci)), p(Gn(Ci, D,A)), i.e. series

$$dG\{G[G(I_i)\ldots]\} = d[G1..f(\Sigma[s(C_i,D)] * \Sigma[p1..f(C_ik,D,A)])]$$

$$= d[\Sigma[s(G1..f(C_i),G1..f(D))]] * \Sigma[p(G1..f(C_{ik}),G1..f(D),G1..f(A))]]$$

The important steps are:

1. Separate “diagnosis function s” from “sampling function p”
2. Invariance of image transformations G

The following parameters are free of choice:

number and sequence of transformations G (spatial sampling)

probability s(Ci,D) => passive/active sampling

Adjustment of sample size D in relation to

the entire size A	=> random sampling
inter - spatial relation	=> stratified sampling

Figure 5

Derivatives of basic sampling procedures in separating the diagnosis function s from the sampling function p.

actual segmentation procedure the sampling can be performed as active and passive sampling.

Active and passive sampling

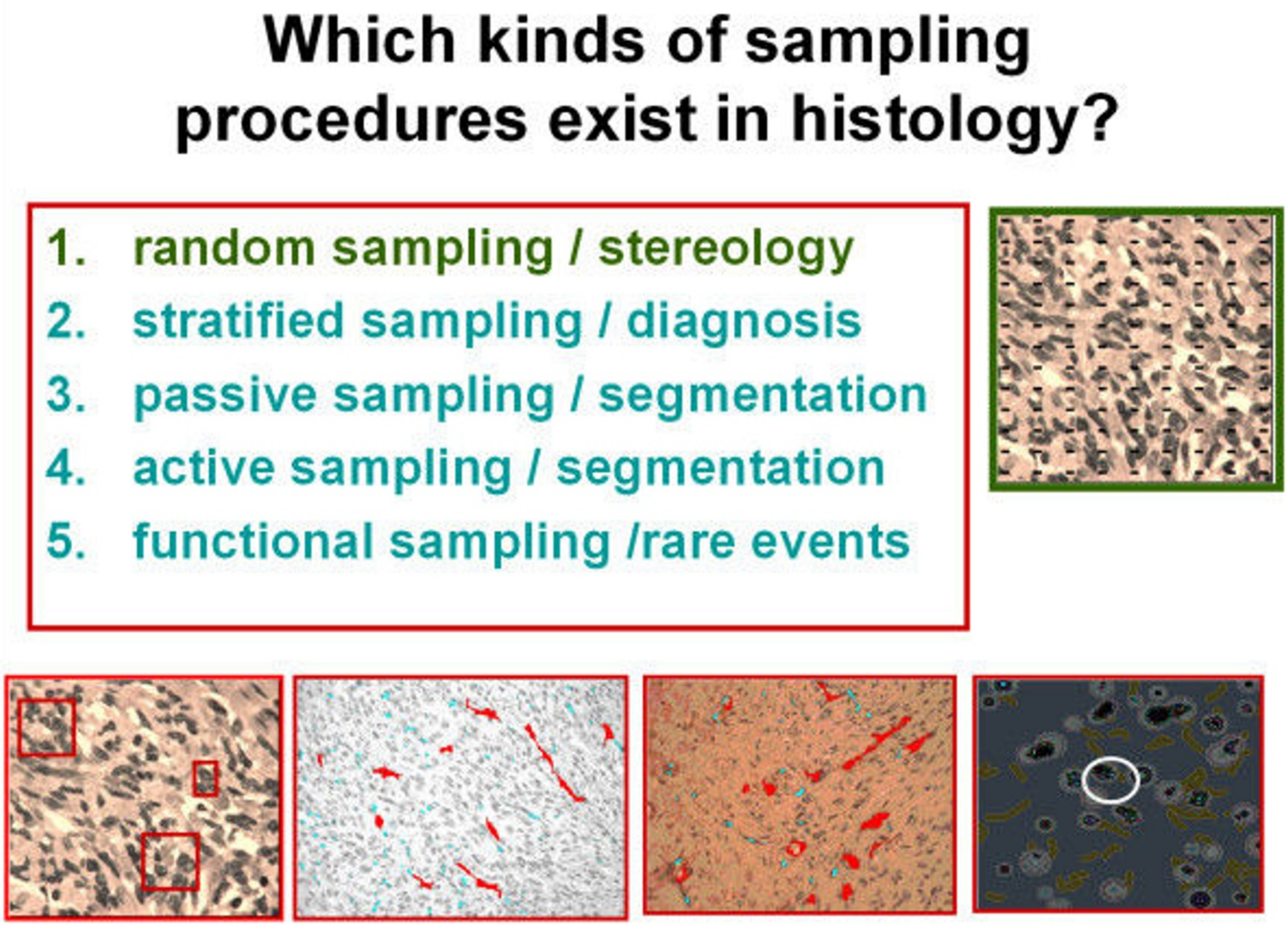
Any segmentation procedure has to accurately define the area of an object, which is equivalent to detect its boundary. Each pixel has to be distinguished either to belong to the object or not, which can be written: f(x, y, meaning) = [1,0], with f(x, y, object) = [1], and f(x, y, backgound) = [0] This approach is called passive sampling, as it discriminates the object area by a simple yes – no function [14]. In other words, passive sampling is provided by a constant relation between the objects and the grid (intersections). The intersection has the probability function p(i) = [1].

Active sampling is a different approach. It is provided by an objective-specific relation between the objects and the grid (intersections). The probability that a pixel belongs to an object ranges between [1,0]. The intersection has a probability function p(i, o), i.e., the probability to detect the pixels that belong to a certain object depends on the object itself and its neighborhood [20]. For example, a pixel displays a probability of 0.7 that it belongs to the object. This probability can increase or decrease dependent upon additional parameters, such as size, orientation, or shape of neighboring objects. Naturally, the probability value of 0.7 itself might be used to define whether it is an "object" – or a "background" pixel.

The probability function p(i, o) can be calculated if we separate p(i, o) in its two components: p(i, o) = gr(x, v) * af(gr, v).

gr(x, y) is the frequency distribution of different objects in the reference space v,

af(gr, v) is the detection probability in the space v.



Which kinds of sampling procedures exist in histology?

1. random sampling / stereology
2. stratified sampling / diagnosis
3. passive sampling / segmentation
4. active sampling / segmentation
5. functional sampling /rare events

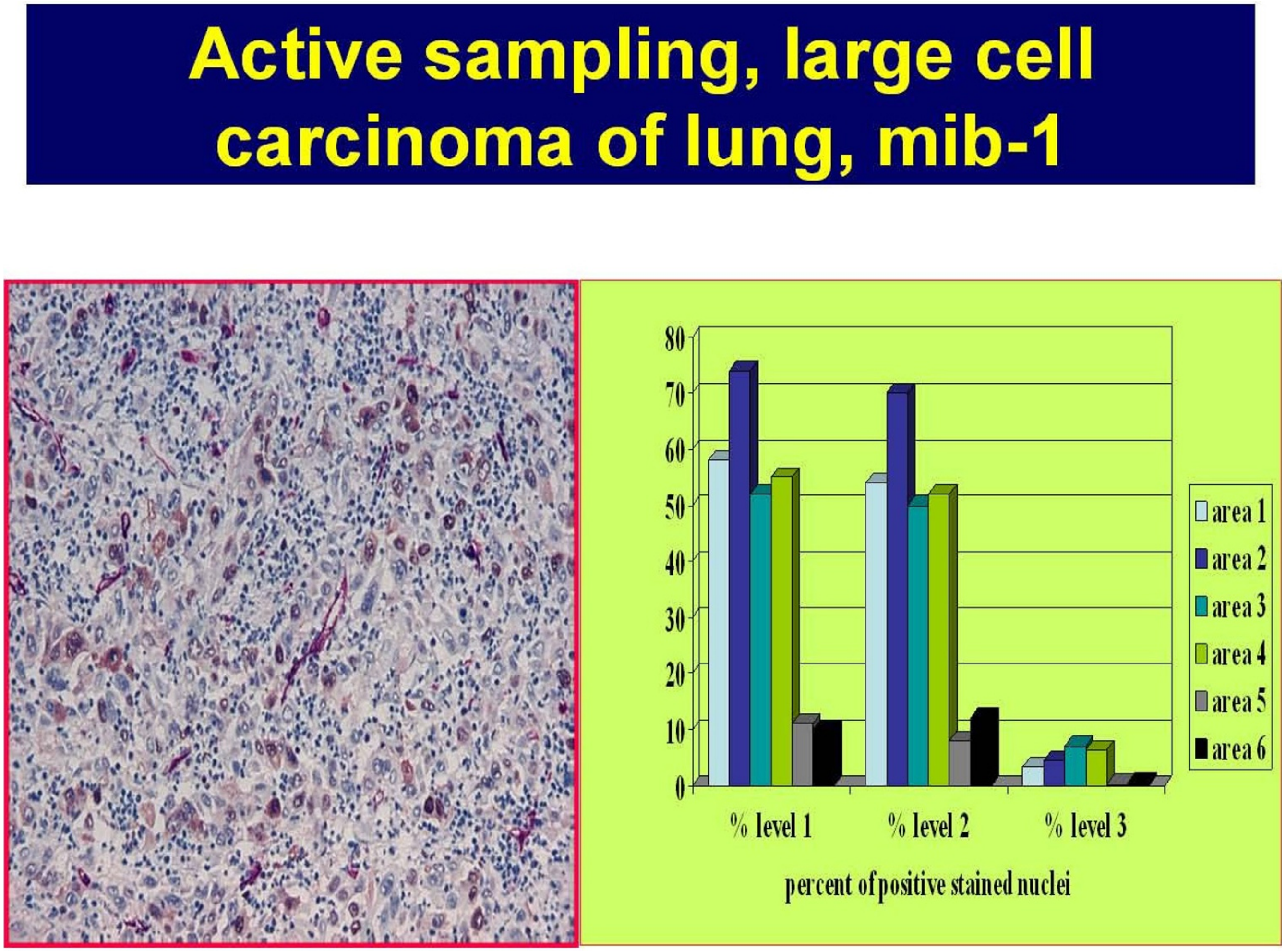
Figure 6
Survey of different sampling procedures.

If we assume that af(gr, v) = const in the reference space v, we can estimate p(i, o) by a set of measurements in different sample spaces and transform p(i, o) = [1] if gr(x, y) > const, and p(I, o) = [0] elsewhere.

Active sampling has been reported to be an effective method to correct the variation in immunohistochemistry staining, for example to compute the threshold of positive staining intensity [2]. The classic problem is: At which staining intensity (color level) can an "immunohistochemically analyzed cell" be grouped into the positive class? The active sampling attempt is to measure the relation positive/negative cells at different threshold levels in several randomly selected sample areas. We then select the discrimination threshold which results in (number of positive objects/number of negative objects) = const for all selected samples. A characteristic application is demonstrated in figure 7. It displays a Mib_1 stain for proliferating nuclei in a lung carcinoma, and the relative number of positive nuclei dependent upon three different thresholds measured in five different sample areas. Threshold number three is obviously too high, and the threshold number 1 too low. The threshold number two should be chosen as discriminating threshold, i.e., to calculate the relative number of proliferating nuclei. This algorithm has been successfully implemented in the automated immunohistochemistry measurement system (EAMUS™) [2,3,9,16,17].

Functional sampling

The idea of functional sampling focuses on the interpretation of rare events [12,14,15,31-33]. The question arises: Do there exist certain rare cells within a cellular society (tissue) that possess a high functional power similar to catalysts in chemistry. If yes, how can they be identified? Therefore, functional sampling is defined as the search for



Active sampling, large cell carcinoma of lung, mib-1

Figure 7

Example of an active sampling to determine the discrimination threshold of a Mib-1 PAP stained undifferentiated large cell carcinoma (in association to the distance of the nearest vessel (proliferating nuclei are stained brown, vessels are stained red). The percentages of stained nuclei in relation to three different thresholds (level 1 – 3) are shown for six different sample areas. The discrimination threshold can be chosen between level 1 and level 2 without major error in contrast to level 3, which would result in wrong (too low) estimates.

a specific (key) function of rare biological objects within a different (majority) population. As the function to be analyzed might be unknown and we cannot observe the proposed function directly, we have to state the following prerequisites for proper analysis:

1. The specific object (cell) is rare within the basic population.

2. It has to possess regular neighborhood relations to objects of the basic population.

3. It has to be randomly distributed within the reference space.

The proposed algorithm tries to evaluate the distance properties between the rare events and the frequent events, and the general distribution of rare events within the reference space as follows:

1. We perform a random sampling of the specific (rare) object (*O*) within the basic population *Ni* (to estimating *O [Ni]*).

2. We perform a stratified sampling "around" each detected specific object (to estimating *Ni(0)*).

3. *If Ni(O) = constant* we can assume a specific function of the object (cell) within the basic population (for example

cellular immune competence, functional activation of cells, etc.).

An example is shown in figure 8 which displays the binding capacities of labeled galectin-1 in an undifferentiated lung carcinoma. Only one large intensively stained cell is present in this sample area (marked by an arrow). Its frequency in all samples is < 5%, and the mean distance between these cells measures 245 ± 198 μm, i.e. these cells do not express a constant distance in between. The opposite, however, holds true for the distance between these large cells and their nearest (majority) cells as well as between the other cells within this tissue. From these figures we can derive, that the rare large cells probably posses a significant biological function for the whole tissue (according to biochemical investigations galectin-1 belongs to a family of galactoside binding proteins that has growth regulatory and immunomodulatory properties [21].

Sampling efficiency

As we have defined sampling, it is a procedure that wants to describe space and time-related properties in surgical pathology, i.e. in tissue – based diagnosis. Such an investigation can be performed in different manners, which can be of different efficiencies. How can sampling efficiency be measured? Obviously, any sampling efficiency is closely related to the spatial distribution of the events searched for in the reference space. If the spatial distribution is known we can adjust the sampling procedure correspondingly. However, its spatial distribution is often not known, and we have to start with a random selection

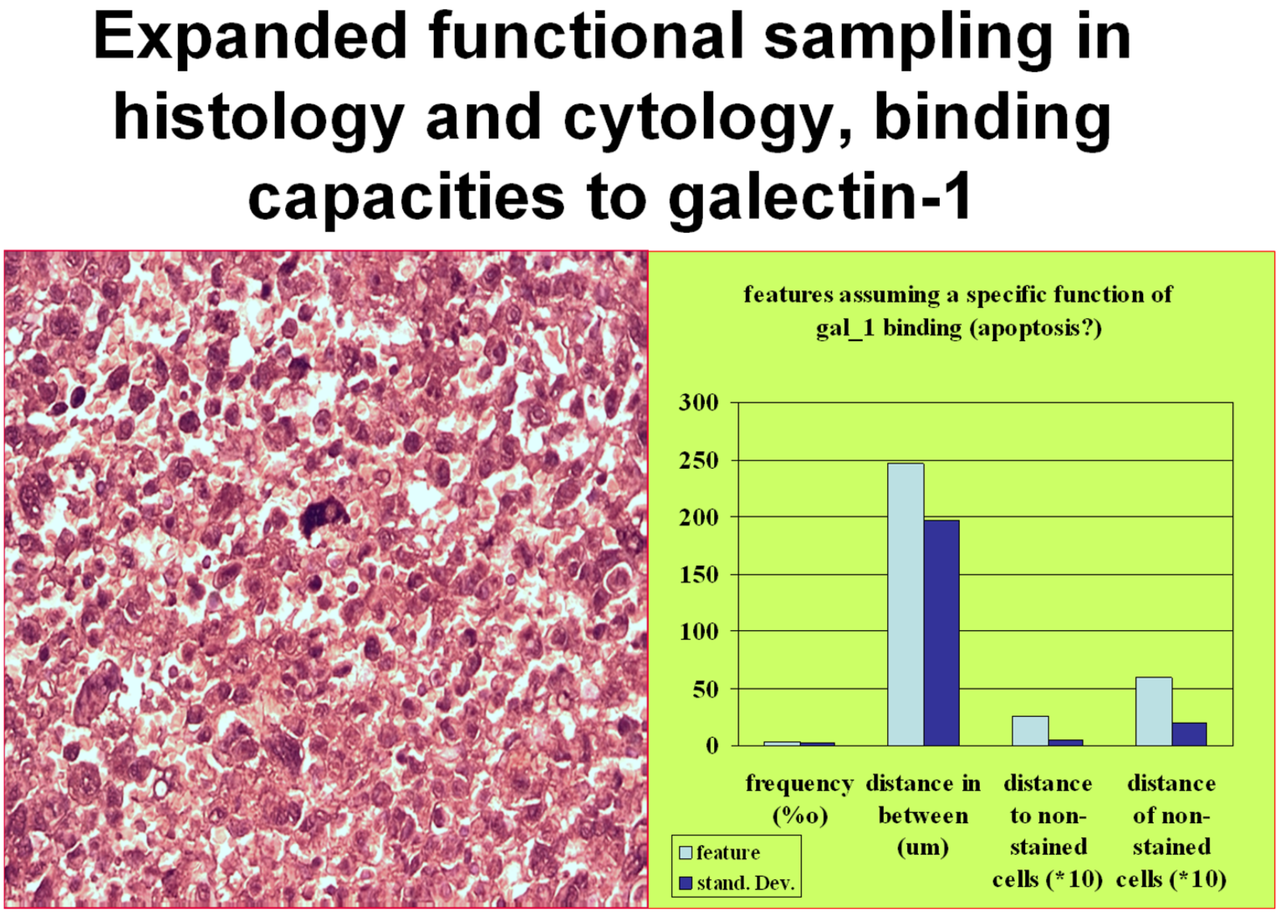


Figure 8

Example of expanded functional sampling: A rare event (galectin-1 binding large cancer cell) within a large cell anaplastic lung carcinoma displays a regular distance to its nearest neighbouring cells and a random distribution within the whole slide. In combination with biochemical data the result supports the idea that these cells might posses a "catalyst" function for progression of the total cancer cell population.

of certain compartments (samples). We can then measure the spatial distribution (frequency) p = s(N)/v, and the variation of p is the error E(p), which should be small in order to have an efficient sampling procedure. Usually, we can state that the reference space v >> s(Ne) (size of element e). The error E(p) can then be calculated according to

$$E(p) = \Sigma[E^2(Ne) + E^2(B(n)) + E^2(Ne/sv)]^{(1/2)}$$

with

E(Ne) = error of detecting an individual event (i.e., probability of identification/missing a tumor cell)

E(B(n)) = error of measuring all elements in the reference space (i.e., related to the biological variance of the tissue, dependent upon N)

E(Ne/v) = error of measuring the size of events e in relation to size of sampling space S (frequency of e in sample space sv).

We can derive the following statements from this formula:

1. We obtain the smallest sampling error if we select the reference volume as sample size, and if we are dealing with regular tissue (small biological variance).

2. The smaller the sample sizes in relation to the size of events, the bigger is the sampling error, as long as the error to segment (identify) per event is not increasing.

3. The sampling error is increasing if we choose different sizes of the samples.

Discussion

To take and to analyze samples of a broad variety of tissues is a basic procedure in surgical pathology, or in tissue – based diagnosis. All diagnostic algorithms depend upon a correct and reliable sampling procedure, and extensive training in surgical pathology addresses to identify and sample those tissue compartments that probably contain the most significant information to classify the disease present [7,19,34-38]. The majority of investigations addresses to an optimum sampling procedure, for example. How many sentinel lymph nodes should be investigated in relation to the stage of breast cancer [29,31], or "optimizing sampling of tomato fruit for carotenoid content, or how to perform endometrial sampling in patients with trophoblastic disease after suction curettage [39,40]. In the early days of stereology several authors took attention on the sampling procedures, as the results of counting interceptions are closely associated to the nature of the used sampling method [22,23]. Recently, sampling has returned to the focus of investigations, especially in live imaging [41]. Most of the investigations try to optimize the sampling, which is equivalent to evaluate the "best" stratified sampling method.

In addition to medical applications, sampling plays a dominant role in geology, especially mining. In fact, Krige's sampling analysis can be considered to be the first approach to develop a "sampling theory" [12,20].

In this article we want to derive a scheme of sampling that permits a principle view of sampling, its different methods, and to calculate the efficiency of the used sampling method. In principle, two different algorithms exist, the random sampling and the stratified sampling [12,9]. Random sampling has to be performed, if no knowledge of the information searched for exists. It is the appropriate technique to measure features of biological units such as chromosomes, DNA fragments, nuclei, cells, vessels, etc. Its accuracy (error rate) can be predefined by number and size of the chosen samples in relation to the expected size of events and to the reference space. Its results can be implemented in additional classification algorithms, such as diagnostic procedures. The sampling can be terminated if a certain classification can be performed with a predefined accuracy, i.e, a diagnosis can be assessed with high certainty. The accurate measurement of events' features is a prerequisite, but not the aim of stratified sampling. Its implementation requires additional (external) information, and numerous investigations have been performed to "speed up" the procedure (or to make it more efficient) using spatial structures within the reference space. When an exponential event probability distribution is given, Krige's formula can be derived from stratified sampling.

In addition to the discussed principle differences between random and stratified sampling procedures, passive and active sampling plays a major role in image segmentation algorithms. The common principle of active sampling associates neighbourhood knowledge (i.e. knowledge derived from general external observations) to the object under investigation, for example to accurately define its boundaries [18]. Especially in measuring accurate thresholds for grading purposes in immunohistochemistry this approach has been proven to be successful [2]. A furthermore derived application is the functional sampling, which is again a stratified sampling in principle. This procedure can assist to investigate in the "biological importance" of rare events, which is widely not known to our experience.

In aggregate, a general theory of sampling is derived that possesses its applications in numerous, if not all natural sciences. They range from agriculture to mining, from aircraft maintenance to medicine. In surgical pathology it is of major importance that all diagnostic investigations start with appropriate sampling.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KK and EV initialized the study and drafted the paper. HS, TG, JG and GK were involved in generating and evaluating the data and in the writing of the manuscript.

Acknowledgements

The financial support of the international Academy of Telepathology e.V., and of the Verein zur Förderung des biologisch-technologischen Fortschritts in der Medizin e.V. are gratefully acknowledged.

References

1. Harewood GC: **Endoscopic tissue diagnosis of cholangiocarcinoma.** *Current opinion in gastroenterology* 2008, **24:**627-630.
2. Kayser G, Radziszowski D, Bzdyl P, *et al.*: **Theory and implementation of an electronic, automated measurement system for images obtained from immunohistochemically stained slides.** *Anal Quant Cytol Histol* 2006, **28:**27-38.
3. Kayser K, Molnar B, Weinstein R: **Virtual Microscopy: Fundamentals, Applications, Perspectives of Electronic Tissue-based Diagnosis.** Berlin: VSV Interdisciplinary Medical Publishing; 2006.
4. Molnar B, Berczi L, Diczhazy C, *et al.*: **Digital slide and virtual microscopy based routine and telepathology evaluation of routine gastrointestinal biopsy specimens.** *J Clin Pathol* 2003, **56:**433-438.
5. Kayser K, Kayser G: **Basic aspects of and recent developments in telepathology in Europe, with specific emphasis on quality assurance.** *Anal Quant Cytol Histol* 1999, **21:**319-328.
6. Kayser K, Kayser G, Radziszowski D, *et al.*: **New developments in digital pathology: from telepathology to virtual pathology laboratory.** *Stud Health Technol Inform* 2004, **105:**61-69.
7. Leong FJ, McGee JO: **Automated complete slide digitization: a medium for simultaneous viewing by multiple pathologists.** *J Pathol* 2001, **195:**508-514.
8. Leong FJ, Nicholson AG, McGee JO: **Robotic telepathology: efficacy and usability in pulmonary pathology.** *J Pathol* 2002, **197:**211-217.
9. Kayser K, Kayser G: **Virtual Microscopy and Automated Diagnosis.** In *Virtual Microscopy and Virtual Slides in Teaching, Diagnosis and Research* Edited by: Gu J, Ogilvie R. Boca Raton: Taylor & Francis; 2005.
10. Florell SR, Smoller BR, Boucher KM, *et al.*: **Sampling of melanocytic nevi for research purposes: a prospective, pilot study to determine effect on diagnosis.** *Journal of the American Academy of Dermatology* 2008, **59:**814-821.
11. Kayser K, Kayser G, Becker HD, *et al.*: **Telediagnosis of transbronchial fine needle aspirations--a feasibility study.** *Anal Cell Pathol* 2000, **21:**207-212.
12. Kayser K, Hufnagl P, Kayser G, *et al.*: **Stratified sampling: Basic ideas and application in pathology.** *Elec J Pathol Histol* 1999, **5:**994-906.
13. Mavrodieva V, Levy L, Gabriel DW: **Improved sampling methods for real-time polymerase chain reaction diagnosis of citrus canker from field samples.** *Phytopathology* 2004, **94:**61-68.
14. Trpkov K, Thompson J, Kulaga A, *et al.*: **How much tissue sampling is required when unsuspected minimal prostate carcinoma is identified on transurethral resection?** *Archives of pathology & laboratory medicine* 2008, **132:**1313-1316.
15. Winters R, Waters BL: **What is adequate sampling of extraplacental membranes?: a randomized, prospective analysis.** *Archives of pathology & laboratory medicine* 2008, **132:**1920-1923.
16. Kayser K, Gortler J, Goldmann T, *et al.*: **Image standards in Tissue-Based Diagnosis (Diagnostic Surgical Pathology).** *Diagn Pathol* 2008, **3:**17.
17. Kayser K, Gortler J, Metze K, *et al.*: **How to measure image quality in tissue-based diagnosis (diagnostic surgical pathology).** *Diagn Pathol* 2008, **3(Suppl 1):**S11.
18. Kayser K, Hoshang SA, Metze K, *et al.*: **Texture and object related automated information analysis in histological still images of various origins.** *Analytical and Quantitative Cytology and Histology* 2008, **30:**323-35.
19. Kayser K, Radziszowski D, Bzdyl P, *et al.*: **Towards an automated virtual slide screening: theoretical considerations and practical experiences of automated tissue-based virtual diagnosis to be implemented in the Internet.** *Diagn Pathol* 2006, **1:**10.
20. Krige D, DS Rand: **Some basic considerations in the application of gepstatistics to the valuation of ore in South African gold mines.** *J South Afr Inst Min Metal* 1976, **77:**383-391.
21. Kayser K, Gabius HJ: **Graph theory and the entropy concept in histochemistry. Theoretical considerations, application in histopathology and the combination with receptor-specific approaches.** *Prog Histochem Cytochem* 1997, **32:**1-106.
22. Gundersen HJ: **Stereology of arbitrary particles. A review of unbiased number and size estimators and the presentation of some new ones, in memory of William R. Thompson.** *Journal of microscopy* 1986, **143:**3-45.
23. Gundersen HJ, Jensen EB: **Stereological estimation of the volume-weighted mean volume of arbitrary particles observed on random sections.** *Journal of microscopy* 1985, **138:**127-142.
24. Knust J, Ochs M, Gundersen HJ, *et al.*: **Stereological estimates of alveolar number and size and capillary length and surface area in mice lungs.** *Anat Rec (Hoboken)* 2009, **292:**113-122.
25. Vesterby A, Gundersen HJ, Melsen F: **Unbiased stereological estimation of osteoid and resorption fractional surfaces in trabecular bone using vertical sections: sampling efficiency and biological variation.** *Bone* 1987, **8:**333-337.
26. Vesterby A, Kragstrup J, Gundersen HJ, *et al.*: **Unbiased stereologic estimation of surface density in bone using vertical sections.** *Bone* 1987, **8:**13-17.
27. Mayhew TM: **A stereological perspective on placental morphology in normal and complicated pregnancies.** *Journal of anatomy* 2008 in press.
28. Mayhew TM, Muhlfeld C, Vanhecke D, *et al.*: **A review of recent methods for efficiently quantifying immunogold and other nanoparticles using TEM sections through cells, tissues and organs.** *Ann Anat* 2008 in press.
29. Michel SC, Low R, Singer G, *et al.*: **[Stereotactic Mammotome breast biopsy: routine clinical experience and correlation with BI-RADS--classification and histopathology].** *Praxis* 2007, **96:**1459-1474.
30. Kennedy MP, Jimenez CA, Bruzzi JF, *et al.*: **Endobronchial ultrasound-guided transbronchial needle aspiration in the diagnosis of lymphoma.** *Thorax* 2008, **63:**360-365.
31. Atula T, Shoaib T, Ross GL, *et al.*: **How many sentinel nodes should be harvested in oral squamous cell carcinoma?** *Eur Arch Otorhinolaryngol* 2008, **265(Suppl 1):**S19-23.
32. Trotz M, Weber MA, Jacques TS, *et al.*: **Disseminated langerhans cell histiocytosis-related sudden unexpected death in infancy.** *Fetal and pediatric pathology* 2009, **28:**39-44.
33. Zong JC, Arav-Boger R, Alcendor DJ, *et al.*: **Reflections on the interpretation of heterogeneity and strain differences based on very limited PCR sequence data from Kaposi's sarcoma-associated herpesvirus genomes.** *J Clin Virol* 2007, **40:**1-8.
34. Bartels P, Weber J, Duckstein L: **Machine learning in quantitative histopathology.** *Anal Quant Cytol Histol* 1988, **10:**299-306.
35. Bartels PH: **Computer-generated diagnosis and image analysis.** *An overview. Cancer* 1992, **69:**1636-1638.
36. Oger M, Belhomme P, Klossa J, *et al.*: **Automated region of interest retrieval and classification using spectral analysis.** *Diagnostic Pathology* 2008, **3(Suppl 1):**S17.
37. Permuth-Wey J, Boulware D, Valkov N, *et al.*: **Sampling strategies for tissue microarrays to evaluate biomarkers in ovarian cancer.** *Cancer Epidemiol Biomarkers Prev* 2009, **18:**28-34.
38. van Diest PJ, Kayser K, Meijer GA, *et al.*: **Syntactic structure analysis.** *Pathologica* 1995, **87:**255-262.
39. Darrigues A, Schwartz SJ, Francis DM: **Optimizing sampling of tomato fruit for carotenoid content with application to assessing the impact of ripening disorders.** *Journal of agricultural and food chemistry* 2008, **56:**483-487.

40. Lertkhachonsuk R, Treratanachat S: **Endometrial sampling in patients with trophoblastic disease after suction curettage.** *The Journal of reproductive medicine* 2008, **53:**634-638.
41. Lundstrom C, Ljung P, Persson A, *et al.*: **Uncertainty visualization in medical volume rendering using probabilistic animation.** *IEEE transactions on visualization and computer graphics* 2007, **13:**1648-1655.